W 1413

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-032290
(43)Date of publication of application : 31.01.2003

(51)Int.Cl. H04L 12/56
G06F 3/06
G06F 12/00
G06F 13/00

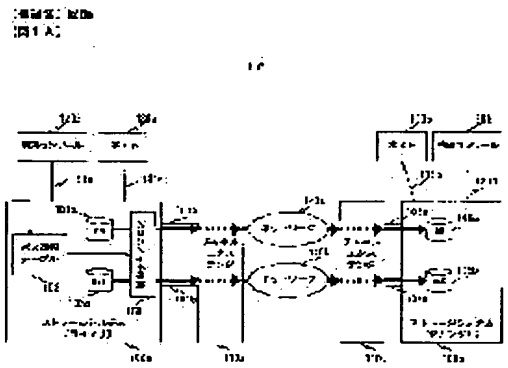
(21)Application number : 2002-082329
(22)Date of filing : 25.03.2002

(71)Applicant : HITACHI LTD
(72)Inventor : YAMAKAMI KENJI

(30)Priority
Priority number : 2001 823470 Priority date : 30.03.2001 Priority country : US

(54) PATH SELECTION METHOD FOR STORAGE BASED REMOTE COPY

(57)Abstract:
PROBLEM TO BE SOLVED: To provide techniques for managing data flow over a plurality of connections between primary and remote storage devices.
SOLUTION: In a representative embodiment, when a primary storage system copies data to a secondary storage system, it chooses one of a plurality of networks connected to the secondary storage system, depending upon a user's policy. Since the networks have different characteristics, in terms of for example, performance, security, reliability and costs, the user can specify which networks are used under various circumstances, i.e., daytime operation, nighttime operation, normal operation, emergency, and so forth. The storage systems comprise a mapping of volumes and ports. When performing copy operations, the primary storage system finds a volume storing the data, and available ports by accessing mappings. The mappings are based upon policies that are inputted by a user.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2003－32290
（P2003－32290A）
(43)公開日　平成15年１月31日(2003.1.31)

(51)Int.Cl.[7]		識別記号	FI		テーマコード(参考)
Ｈ０４Ｌ	12/56	１００	Ｈ０４Ｌ 12/56	１００Ｃ	５Ｂ０６５
Ｇ０６Ｆ	3/06	３０１	Ｇ０６Ｆ 3/06	３０１Ｍ	５Ｂ０８２
				３０１Ｘ	５Ｂ０８９
	12/00	５３１	12/00	５３１Ｄ	５Ｋ０３０
	13/00	３５４	13/00	３５４Ａ	

審査請求　未請求　請求項の数20　ＯＬ　外国語出願　（全 59 頁）

(21)出願番号　特願2002－82329(P2002－82329)

(22)出願日　平成14年３月25日(2002.3.25)

(31)優先権主張番号　０９／８２３４７０

(32)優先日　平成13年３月30日(2001.3.30)

(33)優先権主張国　米国（ＵＳ）

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　山神　憲司
アメリカ合衆国カリフォルニア州　ロスガトス　カレニベル108

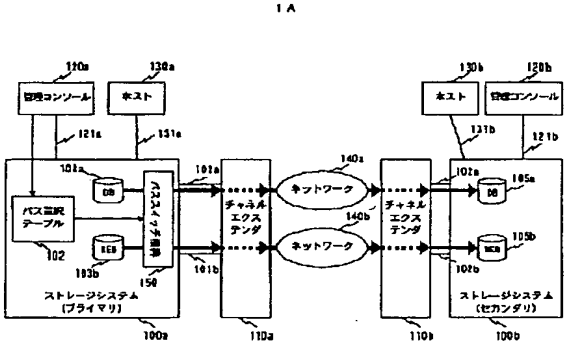
(74)代理人　100075096
弁理士　作田　康夫

Ｆターム(参考)　5B065 BA01 CA02 CE02 CH11
5B082 DA02
5B089 GA21 HB19 LB25
5K030 GA20 HC01 KA02 KA13 LB02
LB05 MB01

(54)【発明の名称】　ストレージベースリモートコピーでのパス選択方式

(57)【要約】
【課題】　本発明は、プライマリとリモートストレージデバイスとの間の複数接続を通して、データの流れを管理する技術を提供する。
【解決手段】本発明の代表的実施例では、プライマリストレージシステムがセカンダリストレージシステムへデータをコピーするのに、セカンダリストレージシステムに接続する複数のネットワークのうちの一つを、ユーザのポリシーに従って選択する。ネットワークは、例えば、性能、安全性、信頼性、及びコスト等の異なった特性を持つ為、ユーザは、例えば、昼間帯、夜間帯、平常時、及び異常時等の使用環境に応じて、使用すべきネットワークを指定することが可能になる。ストレージシステムは、ボリュームとポートの対応マップを有する。コピー操作を行うときには、プライマリストレージシステムは、対象データを保有するボリュームを見つけ、対応マップにアクセスして使用可能なポートを選択する。対応マップは、ユーザにより入力されたポリシーに準拠している。

【書類名】図面
【図１Ａ】



【特許請求の範囲】
【請求項１】複数のデイスクドライブ内の少なくとも1台と、
パス選択情報を記憶するメモリと、
複数のネットワークにスイッチ可能な接続を備えた複数のポートと、
プロセッサで構成されるストレージシステム装置であって、
前記複数のネットワークの各々は、各々に対応して、ユーザから指定された複数のポリシーの少なくとも一つを持ち、
前記プロセッサは、前記複数のネットワークの複数の状態中の少なくとも一つの状態をモニタし、
前記複数のネットワークの複数の状態中の少なくとも一つの状態とユーザから指定の複数のポリシーとを比較することにより、前記複数のネットワークを接続する前記複数のポートの少なくとも一つを選択するストレージシステム装置。
【請求項２】請求項１記載のストレージシステム装置であって、
前記複数の状態中の少なくとも一つの状態は、スループット、使用率、エラー率、及びエラーの存在の少なくとも一つから成ることを特徴とするストレージシステム装置。
【請求項３】請求項１記載のストレージシステム装置であって、
さらに、複数の状態表示を有し、前記複数のネットワークは、各々に対応して、前記複数の状態表示の少なくとも一つを有し、
前記プロセッサが、前記状態表示に基づいて、前記複数のネットワークを接続する複数のポートの少なくとも一つからひとつのポートを選択するかどうかを決定するストレージシステム装置。
【請求項４】請求項３記載のストレージシステム装置であって、
さらに、ネットワークモニタを有し、
前記ネットワークモニタが、前記複数のネットワークの少なくとも一つの状態を検出して、前記状態表示に値をセットするストレージシステム装置。
【請求項５】請求項３記載のストレージシステム装置であって、前記状態表示が、使用可能、一時的使用不能、及び使用不能の少なくとも一つから成るストレージシステム装置。
【請求項６】請求項１記載のストレージシステム装置であって、前記ポリシーが、限界値、最大値、最小値、平均値、平均、限界、制約、優先度、及び目標の少なくとも一つから成るストレージシステム装置。
【請求項７】請求項１記載のストレージシステム装置であって、
前記複数のネットワークが、複数のパスグループにグループ化され、前記ポリシーは一つのパスグループ内のネットワークに対応するストレージシステム装置。
【請求項８】請求項７記載のストレージシステム装置であって、
前記複数のデイスクドライブの少なくとも１台が、複数のボリュームの少なくとも１つから成るストレージシステム装置。
【請求項９】請求項８記載のストレージシステム装置であって、
前記複数のボリュームの少なくとも１つが、前記複数のパスグループ内の少なくとも一つのパスグループのネットワークへのアクセスが許可されているストレージシステム装置。
【請求項１０】ストレージ装置がネットワークを使用する為のコストを最小にする方法であって、前記方法は、
データ転送に使用する第１のネットワークを指定すること、
前記第１のネットワークに対する制約を指定すること、
データ転送に使用する第２のネットワークを指定すること、
前記第１のネットワークの状態が前記制約を満足するときは、前記第１のネットワークを使用してデー　タ転送を行い、そうでなければ、前記第２のネットワークをデータ転送に使用すること、から成る方法。
【請求項１１】請求項１０記載の方法であって、さらに、
前記第１のネットワークの状態が前記制約を満足しなくても、前記第１のネットワークをテスト目的に使って、前期データの一部分を転送すること、
前記テスト使用中に前記第１のネットワークの状態をモニタすること、
前記テスト結果で、前記第１のネットワークの状態が前記制約を再び満足出来るようになれば、前記第１のネットワークを使用してのデータ転送に戻すこと、から成る方法。
【請求項１２】請求項１０記載の方法において、
前記第１のネットワークは、前記第２のネットワークよりも相対的に使用料が安価である方法。
【請求項１３】請求項１０記載の方法において、
前記第１のネットワークに対する前記制約の指定が、スループット、使用率、エラー率、及びエラーの存在の少なくとも一つを指定することから成る方法。
【請求項１４】請求項１０記載の方法において、
前記第１のネットワークが、パブリックネットワークであり、前記第２のネットワークはプライベートネットワークである方法。
【請求項１５】請求項１０記載の方法において、
さらに、前記第１のネットワークに、前記第２のネットワークよりも高いプライオリテイを設定すること、から成る方法。

【請求項１６】請求項１０記載の方法において、
さらに、前記第１のネットワークでの異常状態を検出したら、前記第２のネットワークを使用してデータ転送を行うこと、から成る方法。
【請求項１７】ネットワークを選択する方法であって、前記方法は、
複数のネットワークの複数の状態の少なくとも一つをモニタすること、
前記複数ネットワークの前記複数状態の少なくとも一つを、複数のユーザ指定ポリシーの少なくとも一つと比較すること、
前記複数ネットワークに接続している複数ポート中の少なくとも一つを選択すること、から成り、
前記複数ネットワークの各々は、前記複数のユーザ指定ポリシーの少なくとも一つを持っている方法。
【請求項１８】請求項１７記載の方法において、前記複数ネットワークに接続している複数ポート中の少なくとも一つの選択が、
状態表示をもとに、前記複数ネットワークを接続する前記複数ポート中の少なくとも一つからポートを選択するかどうかを決定することである方法。
【請求項１９】請求項１７記載の方法において、さらに、
前記複数ネットワークを複数のパスグループに対応付けする、ことから成り、
前記複数ポリシーの少なくとも一つは、複数パスグループの少なくとも一つに対応する方法。
【請求項２０】請求項１７記載の方法において、
複数ネットワークの複数状態の少なくとも一つの状態のモニタリングは、
ネットワークモニタを使用して、前記複数ネットワークの少なくとも一つの状態を検出し、この結果に従って、状態表示に値をセットすること、から成る方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、一般的にリモートデータストレージシステムに関するものであり、特に、プライマリとリモートストレージデバイス間の複数接続パス上でのデータフローを管理するための制御技術に関連する。
【０００２】
【従来の技術】情報技術の発展とともに、ビジネス企業体は益々増大するストレージ容量が必要になっている。平均的フォーチュン１０００企業においては、来るべき年には、倍以上のストレージ容量が必要になると予想されている。加えて、容量の増大は熟練した情報技術者の不足をもたらしている。このため、多くの企業は、情報技術への投資の拡大を余儀なくさせられている。データ損失災害を防止するため、ストレージベースのリモートコピーを採用する企業が増加している。
【０００３】リモートコピーは、ストレージボリュームのミラーイメージをローカルまたはプライマリストレージシステムとリモートまたはセカンダリストレージシステムの間に生成して管理する。プライマリ及びセカンダリストレージシステムは互いにかなりの距離を隔てて設置する事が出来る。この２つのデイスクシステムはネットワークで接続され、本ネットワークを経由して、ローカルデイスクシステムへの更新結果がリモートデイスクシステムに複写される。今日では、多数のタイプのネットワークが２つのリモートコピー用ストレージシステムを結合可能である。１つのタイプのネットワークは、例えばT3プライベートネットワークの如く、高速で、高信頼で、安全である反面、相対的には高価である。他のタイプのネットワークは、例えばインターネットの如く、相対的には低速で、安全性に欠けるが、安価である。
【０００４】バンキング、ファイナンス、航空機予約等でのオンライン・トランザクション処理(OLTP)のようなビジネスで非常に重要なアプリケーションでは、高速レスポンス、高セキュリテイ、高信頼のリモートコピーが必要となる。他のタイプのアプリケーション、例えば、WEBミラーリング、データウエアハウス、データセンタコンソリデーション、大量データ転送等においては、一般的には、リアルタイムでのコピー機能を必要としないため、上記のような要求は不要である。
【０００５】現状のリモートコピー技術でも、確かにある種の優位性はすでに認識されているが、更なる改良の余地が存在する。例えば、従来からのリモートコピー技術による限り、通信事業会社は、顧客が要求するスループットベースで課金するか、プライベートネットワークでは、サービス単位での支払いを提案する事も有る。例えば、通信事業会社は月単位でのネットワーク帯域幅ベースで課金する事もあり得る。
【０００６】
【発明が解決しようとする課題】しかしながら、リモートコピーユーザによっては、運用上の制限を甘受してでも、データ接続にからむコストの低減を望むと考えられる。例えば、ユーザはアプリケーションの特性に応じて、異なるネットワークを使用して、費用を低減する事が可能である。例えば、WEBミラーリングアプリケーションではインターネットを利用して、バンキングのOLTPではT3ネットワークを使用しても良いはずである。バックアップ目的でストレージを利用したいユーザには常時接続は必要としない。しかしながら、従来の技術では、通信事業者によるデータ転送サービスをアクセス対応の料金で管理する事は不可能である。更に、セキュリテイの如き問題は、ユーザ、通信事業者双方にとって重要な関心事である。
【０００７】この事は、ユーザにとっては、貴重な企業情報資産がリモートストレージに転送されるデータにアクセス制限する事により、保護されるべきことを意味す

る。この事は通信事業者にとって、データの正当性が自らの顧客に対して保持され、通信事業者により認証されない如何なるユーザもアクセスできないことを意味する。

【０００８】真に要求される技術は、プライマリとリモートストレージデバイス間の複数接続上のデータフローを管理する改良された技術である。本発明は、プライマリとリモートストレージデバイスとの間の複数接続でのデータの流れを管理する技術を提供する。

【０００９】

【課題を解決するための手段】本発明の代表的な実施例では、プライマリストレージシステムがセカンダリストレージシステムへデータをコピーするのに、該セカンダリストレージシステムに接続する複数のネットワーク接続の中より１つをユーザのポリシーに従って選択する。ネットワークは、例えば、性能、安全性、信頼性、及びコスト等で異なった特性を持つ為、ユーザは、例えば、昼間帯、夜間帯、平常時、及び異常時等の使用環境に応じて、使用すべきネットワークを指定することが可能になる。ストレージシステムはボリュームとポートの対応マップを有する。プライマリストレージシステムがコピー操作を行うときには、対象データを保有するボリュームを見つけ、該対応マップにアクセスして、使用可能なポートを選択する。対応マップはユーザにより入力されたポリシーに準拠している。

【００１０】１つの代表的な実施例では、プライマリストレージシステムは、ある特定のネットワークを使用するのに、事前設定された最大スループットの範囲内で使用するように制約することが出来る。例えば、ユーザが5MB/sの最大スループットをあるネットワークに指定(制約)すると、プライマリストレージシステムは、その範囲内でしか当該ネットワークを使用しない。5MB/sの限界値に達すると、プライマリストレージシステムは、他のネットワークに接続されたポートを選択する。本メカニズムは、過負荷に脆弱なネットワークがストレージ転送に使用されている場合は、特に顕著な性能向上をもたらす。

【００１１】更に他の実施例では、ユーザが、あるネットワークを使用ベースの支払いで契約している場合は、当該ネットワークを、例えば5MB/s以下の範囲内で使用する事により、余分な出費を抑えることが出来る。更に他の実施例では、プライマリストレージシステムは、安価なパブリックネットワークを、例えばトラフイックが込み合う昼間帯を避けて、使用するようにする事も出来る。更に、プライマリストレージシステムが、ネットワークを通して、リモートコピーデータを送信する事により、該ネットワークの負荷を増大させ、該ネットワークサービスに頼っている他のユーザに悪影響を与えることもあり得る。従って、プライマリストレージシステムはこのような影響を避ける為に、昼間帯は高価ではあるが、プライベートネットワークを使用するようにする事も出来る。

【００１２】更に他の実施例では、プライマリに使用している安価なネットワークが高トラフイック状態になると、該ネットワークを監視している外部接続のネットワークモニタがプライマリストレージシステムに通知する。これにより、プライマリストレージシステムは、外部接続のネットワークモニタが、該プライマリネットワークが平常の低いトラフイックに戻ったことを知らせてくる迄の間は、他のネットワークに切り替えて動作を続ける。他の実施例として、プライマリストレージシステムは、自分がセカンダリストレージシステムに転送すべき未転送データを過剰に保有している状況を検出する。一般的には、安価なプライマリネットワークは高価なセカンダリネットワークよりは低速である。従って、安価なプライマリネットワークのデータ転送速度がプライマリストレージシステムでのデータ転送要求に追いつけない場合は、セカンダリストレージシステムに転送すべきデータがプライマリストレージシステム中で集積してしまうことになる。この状態を放置すると、いずれは、セカンダリストレージシステムはプライマリストレージシステムのミラーイメージコピーを保持する事は不可能になってしまう。この事態を避ける為に、プライマリストレージシステムは自分の中に溜まっている未転送データ量を監視して、一定の限界を超えると、より高価なセカンダリネットワークに切り替える。

【００１３】更に本実施例では、ストレージ装置でのネットワーク使用コストを最小にする方法を提供する。本方法では、まず、データ転送に使用する第１のネットワークを指定する。本方法では、この第１のネットワークに制約を指定する。各種の状況のもとで、制約には、例えば、スループット、使用率、エラー率、及びエラーの存在等からの少なくとも１つが採用されるが、実施例は多様であり、他のタイプの制約が選択されても良い。本方法は、更に、第２のネットワークを指定する事を含む。第１のネットワークが指定された制約条件を満たしている限り、第１のネットワークを使用し続けるが、満たされなくなると、第２のネットワークを使用する事が本方法に含まれる。

【００１４】実施例次第では、第１のネットワークに課せられた制約が満たされない状態でも、テスト目的に一部のデータを第１のネットワークを使用して転送することが本方法に含まれる。本テスト転送により、第１のネットワークの制約条件の回復が確認できたら、再び第１のネットワークの継続的使用に復帰する事も本発明に含まれる。本実施例では、第１のネットワークは第２のネットワークに比べて比較的安価で、又、第１のネットワークはパブリックネットワークで第２のネットワークはプライベートネットワークである。ユーザがネットワークと制約を指定する場合、第１のネットワークを第２の

ネットワークより高いプライオリテイにしても、又は、第１のネットワークでの異常事態の発生を契機に第２のネットワークに切りかえるようにしても良い。

【００１５】他の戦略として、エラー数又はエラー率をモニタする。プライマリストレージシステムはデータ転送中のエラー数をモニタして、エラー率を計算する。もし、エラー率がある限界値を越えて過大になって来たら、プライマリストレージシステムは高価なネットワークに切りかえる。本限界値は、例えば、ユーザのポリシーで決めてもよい。高価なネットワークの使用中でも、プライマリストレージシステムは安価なネットワークをテスト目的に引き続き試用する。安価なネットワークのエラー率が限界値以下に低減すると、プライマリストレージシステムは高価なネットワークの使用を停止する。安価なネットワークにTCP/IPプロトコルが使用されている場合は、高いエラー率は高いトラフィックのためであることが多い。

【００１６】異常事態の発生で、安価なネットワークから、代替としての高価なネットワークへ切り替えるのも本発明の戦略である。この戦略では、プライマリストレージシステムは、平常時は安価なネットワークを使用し、障害発生時にはより高価なネットワークに切りかえる。

【００１７】他の実施例では、ネットワークを選択する方法が提供される。本方法は、複数ネットワークの１つ以上の状態をモニタする事を含む。複数ネットワークの１つ以上の状態を、ユーザから指定された１つ以上のポリシーと比較し、複数ネットワークに接続する１つ以上のポートを選択する事は、本方法の一部である。１実施例では、複数ネットワークの１つ以上の状態をモニタする為に、ネットワークモニタを使用して、１つ以上のネットワークの状態を監視させ、この結果に従って状態表示に該当する値をセットし、又、複数ネットワークに接続する１つ以上の使用ポートを選択するために、該状態表示を利用することが含まれる。複数ネットワークの各々は、ユーザから指定されたポリシーを持っている。本実施例には、複数のネットワークを複数のパスグループに対応させ、１つ以上のパスグループに対して、１つ以上のポリシーを対応させることを含む。

【００１８】更なる実施例では、ストレージ装置が提供される。本ストレージ装置は、１つ以上のデイスクドライブ、パス選択情報を保有するメモリ、複数のネットワークに対して選択的接続する複数のポート、及びプロセッサを有する。複数のネットワークの各々は、ユーザから指定されたポリシーを持っている。代表的なポリシーとしては、例えば、限界値、最大値、最小値、平均値、平均、範囲、制約、優先度、及び目標等が考えられる。プロセッサは、複数ネットワークの１つ以上の状態をモニタして、この結果をユーザ指定のポリシーと比較して、該複数ネットワークに結合する１以上の使用ポートを選択する。代表的な状態としては、例えば、スループット、使用率、エラー率、及びエラーの存在等がある。

【００１９】更に、本実施例では、ストレージ装置は複数の状態表示を持ち、この各々はネットワークに対応している。プロセッサは、本状態表示結果により、複数のネットワークに接続するポートの選択を行う。代表的な状態としては、例えば、”使用可能”、“一時的使用不能”、及び”使用不能”が含まれる。実施例では、１つ以上のネットワークの状態を検出して、状態表示に値をセットするネットワークモニタが提供される。また、実施例によっては、ネットワークは複数のパスグループにグループ化され、ポリシーは１グループ内のネットワークに対応付けることが出来る。更にまた、デイスクドライブはボリュームに分割され、各ボリュームが１つ以上のパスグループ内のネットワークにアクセスできる。

【００２０】本発明は第１と第２のネットワークで構成される例に基づいて記述するが、これは単に説明目的の為であって、本発明の多様な実施例を制限するものではないことに注意が必要である。各種の便益を本明細書で説明する。本発明の更なる本質と便益は、本明細書のこれから後の部分と添付図面を参照することにより明らかになる。

【００２１】

【発明の実施の形態】本発明は、プライマリ・ストレージ機器とリモート・ストレージ機器間の複数接続データフロー管理の改良された技術を提供する。リモートコピー技術は、デイスク・システムの１セットのイメージコピーのミラーを、対となる他のシステムに用意する。その２つのディスクシステムは、ポートにより相互接続しており、互いにある程度離れて設置される。リモートコピーシステムは、ローカル又はプライマリシステムのミラーイメージを保持する。ミラーイメージは、リモート又はセカンダリデイスクシステムに保持される。

【００２２】ローカルデイスクシステムは、ペアのローカルデイスクシステム側にデータをコピーする。ホストがローカルデイスクシステムのデータを更新すると、ローカルデイスクシステムは、該データのコピーをリモートシステムにポートとネットワークリンクを通じて転送する。従って、ホスト動作としては、ローカルデイスクシステムボリュームのミラーイメージを保持する必要はない。代表的なリモートコピー技術の詳細に付いては、U.S.Patent No.5,459,857及び5,544,347等を参照すると良い。

【００２３】ローカル及びリモートデイスク間のデータ転送に付いては、多様なタイプが存在する。“同期モード”では、ローカルデイスクシステムは、ホストからのWrite要求の完了報告をする前に、リモートデイスクシステムにデータ転送を行う。“準同期モード”では、ローカルデイスクシステムは、ホストからのWrite要求の完了報告をしてから、リモートデイスクシステムにデー

タ転送を行う。どちらのモードでも、前のデータ転送の終了がホストへ示されるまでは、ホストより次のWrite要求の処理は開始されない。

【００２４】これに対して“適応モード”では、リモートデイスクシステムに転送されていないデータは、プライマリデイスクシステムのメモリに保存され、ローカルデイスクシステム及びポートがコピータスクに使用可能になった時点で、リモートデイスクシステムに転送される。従って、ホストシステムによるデイスク書きこみ動作は、リモートストレージシステムへのコピー動作の完了を待つことによる中断無しに継続される。リモートコピーシステムでの代表的転送モードに付いては、U.S.Patent No.5,933,653を参照のこと。

【００２５】図１Ａと図1Ｂは、本発明の１実施例での代表的なシステム構成を示す図である。図１Aに示される如く、プライマリストレージシステム１００ａ及びセカンダリストレージシステム１００ｂと称される２つのストレージシステムは、リモートストレージバックアップシステムを構成する。プライマリストレージシステム１００ａ及びセカンダリストレージシステム１００ｂの各々は、１つ以上のデータ記憶用ボリュームを有する。ストレージシステム１００ａと１００ｂは、プログラム実行の為のプロセッサ及び制御データやテーブルを記憶する為のメモリを有する。

【００２６】動作中は、プライマリストレージシステム１００ａに格納されたデータは、セカンダリストレージシステム１００ｂ内の同一タイプのボリュームにコピーされる。本動作は、ミラーリングとか、ミラーイメージングとか呼ばれることもある。例えば、プライマリストレージシステム１００ａ内のボリューム１０３ａや１０３ｂに格納された情報は、セカンダリストレージシステム１００ｂ内のボリューム１０５ａや１０５ｂにミラーされる。プライマリストレージシステム１００ａ及びセカンダリストレージシステム１００ｂは、１つのエンティティ下にあるかもしれないし、代わりに、サービスプロバイダがプライマリストレージシステム１００ａの所有者にバックアップサービスを提供する為にストレージシステムを準備することがあるかもしれない。

【００２７】加えて、実施例によっては、プライマリコピーとセカンダリ又はバックアップコピーの役割が逆転していたり、更には、この２つのシステムの間で共有されることもあり得る。このような実施例では、セカンダリストレージシステム１００ｂがプライマリストレージシステム１００ａの幾つかのボリュームのミラー役を果たしたり、逆にプライマリストレージシステム１００ａがセカンダリストレージシステム１００ｂの幾つかのボリュームのミラー役を果たすこともある。

【００２８】ホスト１３０ａや１３０ｂの如き１つ以上のホストシステムが、チャネルパス１３１ａや１３１ｂを通して少なくとも一つのプライマリストレージシステム１００ａやセカンダリストレージシステム１００ｂに接続する。ある実施例においては、チャネルパス１３１ａや１３１ｂにSCSI、ファイバチャネル、ESCON等が使用される。ホストシステム１３０ａと１３０ｂは、チャネルパス１３１ａや１３１ｂを通して、プライマリストレージシステム１００ａやセカンダリストレージシステム１００ｂのボリューム１０３ａと１０３ｂ、や１０５ａと１０５ｂ上に保存されたデータにそれぞれアクセスする。

【００２９】管理コンソール１２０ａや１２０ｂは、それぞれがパス１２１ａや１２１ｂを通して、プライマリストレージシステム１００ａやセカンダリストレージシステム１００ｂに接続する。実施例によっては、パス１２１ａや１２１ｂは、LAN、専用パス、SCSI、ファイバチャネル、ESCON等であり得る。管理者は、管理コンソール１２０ａを通して、パス選択テーブルの生成ポリシーを入力する。

【００３０】代表的な実施例では、ネットワーク１４０ａは、パブリック（公共のもの）で、低パフォーマンスで、低セキュリテイだが、比較的低コストのネットワークである。ある実施例においては、ネットワーク１４０ａはインターネットである。ここで、”パブリック(公共のもの)”という用語は、認証された人（時には認証されていないときもある）なら実質的には誰でもアクセス可能なネットワークを意味する。ネットワーク１４０ｂは、プライベートで、高パフォーマンスで、高セキュリテイだが、比較的高コストのネットワークである。ある実施例では、ネットワーク１４０ｂはT3通信回線である。ここで、”プライベート”なネットワークは、１ユーザまたはユーザグループに専有され、他のユーザはアクセス不能なネットワークを意味する。

【００３１】即ち、”パブリック”ネットワークは、他のすべてのネットワークを示す。本発明は、簡明の為に、プライマリストレージシステム１００ａとセカンダリストレージシステム１００ｂの間を、２種類のみのネットワークで接続する単純な実施例で説明する。しかしながら、この単純な構成は、説明をわかりやすくするためであり、本発明をこれに限定するものではないことに注意する必要がある。実際の多くの実施例においては、３種類以上の異なったタイプのネットワークが、ここで説明されたと同様な方法で使用されている。

【００３２】図１Aにおいては、複数のチャネルエクステンダ１１０ａや１１０ｂが、ポート１０１ａや１０１ｂとネットワーク１４０ａや１４０ｂ間のプロトコル変換器の役割を果たす。例えば、ポート１０１ａがSCSIで、ネットワーク１４０ａがインターネットの場合は、チャネルエクステンダ１１０ａや１０１ｂが、SCSIフォーマットのデータをTCP/IPプロトコル形式に変換、又その逆を行う。１つ以上のポート１０１ａや１０１ｂがプライマリストレージシステム１００ａとチャネルエクス

テンダ１１０ａを接続する。チャネルエクステンダ１１０ａは、ネットワーク１４０ａや１４０ｂに対する結合役を果たす。

【００３３】ポート１０１ａがネットワーク１４０ａに、ポート１０１ｂがネットワーク１４０ｂに、チャネルエクステンダ１１０ａを通して結合する。１つ以上のポート１０２ａや１０２ｂも、セカンダリストレージシステム１００ｂとチャネルエクステンダ１１０ｂを結合する。チャネルエクステンダ１１０ｂがネットワーク１４０ａや１４０ｂへの接続を可能にする。

【００３４】図１Ｂは、プライマリストレージシステム１００ａとセカンダリストレージシステム１００ｂの両方が、またはいずれか一方が多種類のプロトコルをサポートしている実施例を示す。従って、各々のチャネルエクステンダ１１０ａや１１０ｂは必要ない。この実施例においては、プライマリストレージシステム１００ａ内のポート１０１ａや１０１ｂは、ネットワーク１４０ａや１４０ｂに１つ以上のインターフェース、例えばファイバインターフェース１６０ａやIPインターフェース１７０ａを使用して、直接接続される。同様に、セカンダリストレージシステム１００ｂ内のポート１０２ａや１０２ｂは、それぞれネットワーク１４０ａや１４０ｂにファイバインターフェース１６０ｂやIPインターフェース１７０ｂを介して、直接接続する。

【００３５】図２は、本発明の１実施例でのパスとボリュームの代表的な関係を示す。図２に示される如く、プライマリストレージシステム１００ａ内のボリューム１０３ａ、１０３ｂ、．．．、１０３ｎは、パスグループ２２０ａ，２２０ｂ、．．．，２２０ｍの如き、１つ以上のパスグループを経由して、データを送信できる。各パスグループは、例えば、ネットワーク１４０ａの如きネットワークに接続する為のポート１０１ａや１０１ｂの如き１つ以上のポートを有する。図２においては、パスグループ２２０ａはネットワーク１４０ａに接続するポート１０１ａを有し、パスグループ２２０ｂはネットワーク１４０ｂに接続するポート１０１ｂを有する。

【００３６】パス使用ポリシー２１０は、１つのボリュームを１つ以上のパスグループに対応させ、データをセカンダリストレージシステムに転送するときに使用するパスのプライオリテイを定義する。例えば、図2において、プライマリストレージシステム１００ａがボリューム１０３ｂ内のデータをセカンダリストレージシステム１００ｂ内のボリュームに転送するときには、第一にパスグループ２２０ａのポート１０１ａを選択する。しかしながら、何らかの原因で、パスグループ２２０ａが使用不能の場合は、パスグループ２２０ｂが選択される。もし、パスグループ２２０ｂが使用不能なら、パスグループ２２０ｍが選択される。

【００３７】図2におけるもう一つの例として、ボリューム１０３ａはパスグループ２２０ａのみの使用が許される。本実施例では、パス使用ポリシー２１０は、以下に図３及び図４にて説明するようなテーブルを使用して実現される。

【００３８】図３は、本発明の１実施例での代表的なパス選択テーブルを図示する。図３に示した如く、パス選択テーブル３００は、例えば、図１のボリューム１０３ａや１０３ｂの如きボリュームに対応するボリューム番号３１０を、図２のパスグループ２２０ａや２２０ｂの如きパスグループに対応する１つ以上のパスグループ番号３２０ａや３２０ｂに対応付ける。ボリューム番号３１０は、ストレージシステム中で固有の番号である。プライマリストレージシステム１００ａは、パス選択テーブル３００において、例えば、ボリューム１０３ａや１０３ｂ等の各々に対応する固有のボリューム番号３１０を持つ。

【００３９】パスグループ番号３２０ａや３２０ｂは、本ストレージシステムで定義された各パスグループに対して固有のものである。例えば、３２０ａや３２０ｂの如く、パスグループに対応して２個以上のエントリが存在するなら、本エントリの数は、本ストレージシステムに接続される異なったネットワークの数に対応する。先行するパスグループ番号３２０ａは、後続するパスグループ番号３２０ｂより高いプライオリテイを持つ。プライマリストレージシステム１００ａは、セカンダリストレージシステム１００ｂにデータ転送を行うときには、高いプライオリテイを持つパスグループを第一に選択する。

【００４０】例えば、図３においては、パスグループ番号３２０ａは、パスグループ番号３２０ｂより高いプライオリテイを持つ。テーブルの第２の行において、番号１のボリュームよりデータ転送を行うときには、プライマリストレージシステム１００ａは番号０のパスグループでは無く、番号１のパスグループのポートを第一に選択する。

【００４１】パス選択テーブル３００のパスグループエントリ数より少ないパスグループしかない場合は、本テーブルの余ったエントリには、NULLが記入される。例えば、ボリューム番号０は、パスグループ番号０のみが許され、他の何れのグループも使用できないため、パスグループ番号３２０ｂの対応欄にはNULLが記入されている。

【００４２】図４は、本発明の１実施例での代表的なパスグループテーブルを図示する。図４に示されるように、パスグループテーブル４００は、パス使用ポリシー２１０を使用したパスについてのインフォメーションを示し、パスグループとポートとを対応付ける。パスグループテーブル４００は、１つのストレージシステム内で各パスグループに対して固有に割り当てられたパスグループ番号４１０を有する。制約４２０は、パスグループ内の各パスを使用する上での制約条件を示す。例えば、

パスグループ番号０は5MB/s以下の制約を課されており、データ転送に5MB/sを越えることは出来ない。１つ以上の制約を１つのパスグループに登録する事が出来る。

【００４３】本発明によれば、多様な実施形態に対応して、多様な制約を用いることができる。代表的な制約の例を以下に述べる。ポート番号４３０ａ、４３０ｂ、及び４３０ｃは、各パスグループにおけるポート番号に対応する番号を保持する。各ポートはストレージシステム１００中で固有の番号を持っている。パスグループテーブル４００のポート番号４３０のエントリ欄数より少ないポート数しか持たないパスグループについては、空きエントリにはNULLが表示されている。例えば、パスグループ番号０は、番号０と１の２つのポートしかないため、ポート番号４３０ｃにはNULLが表示されている。

【００４４】ステータス４４０は、当該パスグループの各時点の状態を示す。実施例では、本ステータスは、“使用可能”、“使用不能”、及び“一時的使用不能”を表示する。“使用可能”状態は、プライマリストレージシステム１００ａはパスグループ内の該当ポートを使用できることを示す。“使用不能”状態は、プライマリストレージシステム１００ａは該当ポートを使用できないことを示す。“一時的使用不能”状態は、プライマリストレージシステム１００ａは該当ポートを一定間隔、例えば１分間に１回テスト目的で使用できることを示す。例えば、制約４２０が“エラー率５％未満（Error Rate less than 5%）”で、“一時的使用不能”状態の場合は、パスグループ２２０ａを使用して、例えば１分間に１回データ転送が出来る。プライマリストレージシステム１００ａは結果を監視して、エラー率が５％以下に低下すれば、ステータスを“使用可能”に変更する。

【００４５】図５は、本発明の１実施例での代表的なパス選択プロセスを示すフローチャートである。図５に示される通り、セカンダリストレージシステム１００ｂへのデータ転送が必要になると、プライマリストレージシステム１００ａは複数のステップを実行する。ステップ５００にて、プライマリストレージシステム１００ａは、パス選択テーブル３００にアクセスして、セカンダリストレージシステム１００ｂへのデータ転送の為に使用するパスグループを選択する。プライマリストレージシステム１００ａは、転送すべきデータを保有するボリューム名を知っている為、パス選択テーブル３００の行番号を決定出来る。そして、１巡目はパスグループ番号３２０ａを選択する。ステップ５２０にて、パスグループ２２０ａが制約を満足しない場合は、プライマリストレージシステム１００ａは次ぎの繰り返しで、パスグループ番号３２０ｂを選択する。

【００４６】ステップ５２０では、パスグループがすべての制約を満たすかを調べる。パス選択テーブル３００に登録されているどのパスグループも制約を満たさない場合は、処理はステップ５４０に進む。ステップ５４０において、プライマリストレージシステム１００ａは当該ボリュームのセカンダリストレージシステム１００ｂとの間のミラー動作を停止して、ユーザに警告をレポートする。

【００４７】ステップ５２０にて、制約を満たすパスグループが存在した場合は、ステップ５６０にて、当該パスグループの全てのポートが使用中か否かをチェックする。プライマリストレージシステム１００ａが当該ポートを使用してデータ転送をしている間は、当該ポートは使用中となる。使用中でないポートが存在すれば、ステップ５３０にてプライマリストレージシステム１００ａはその使用されていないポートを選択し、そのポートを介しデータ転送を行う。次に、ステップ５５０にて、データ転送が成功裏に完了したかを判定する。データ転送が成功裏に終了したら、処理は終了する。そうでなければ、ステップ５００に戻って、プライマリストレージシステム１００ａは他のパスグループを調べる。

【００４８】制約

本発明の実施では、多様なタイプの制約を用いることが出来る。以下は各種の実施例で採用できる代表的な制約の例である。以下のリストは制約を完全に網羅しようというものではなく、本発明の多様な実施例で実施できる多くの異なったタイプの制約の幾つかを例示しようというものである。時間制約は、プライマリストレージシステム１００ａが１つのパスグループの使用出来る時刻範囲を制限する。例えば、あるパスグループについて、”9:00pmから 6:00amの間のみ”と時間制限があった場合、現時刻が8:00amなら、プライマリストレージシステム１００ａは当該パスグループを使用できない。プライマリストレージシステム１００ａはタイムクロックを内蔵しており、あるパスグループについて、時間制約が存在する場合、現時刻が制約範囲内か否かを判定するのに使用される。

【００４９】プライマリストレージシステム１００ａは時刻を定期的に(例えば、１分に１回)チェックする。プライマリストレージシステム１００ａは、時間制約が満たされれば、ステータス４４０を“使用可能”にかえる。同様に、時間制約が満たされなければ、ステータス４４０を“使用不能”に設定する。スループット制約は、プライマリストレージシステム１００ａが１つのパスグループについて使用出来る最大スループットを制限する。例えば、“5MB/s”のスループット制約が、あるパスグループに対して課されており、現在のモニタ結果が5.3MB/sであれば、プライマリストレージシステム１００ａは本パスグループ内のパスを使用する事は出来ない。

【００５０】実施例次第で、プライマリストレージシステム１００ａ内のプロセッサ、ハードウエア、及び/又は、ソフトウエア機構が各ポートのスループットを監視

する。例えば、プロセッサが一定時間中、例えば１秒間にあるポートを通って転送されるデータ量をモニタする。その後、各プロセッサで測定されたデータ量が加算され、全体の合計が当該ポートが所属するパスグループのスループットになる。プライマリストレージシステム１００ａは、スループット制約が満たされれば、ステータス４４０を“使用可能”に、スループット制約が満たされなければ、ステータス４４０を“一時的使用不能”に設定する。

【００５１】プライマリストレージシステム１００ａはスループット監視を継続して、スループットが制約以下に低下すれば、ステータス４４０を“使用可能”に変更する。あるパスグループのステータス４４０が“一時的使用不能”を示し続けている間は、プライマリストレージシステム１００ａは、当該パスグループをテスト使用する為に規則的間隔で選択し、テスト目的以外のデータ転送では、他のパスグループを使用する。使用率制約は、プライマリストレージシステム１００ａが１つのパスグループについて使用出来る最大使用率を制限する。ここで、“使用率”とは、１つのネットワーク回線がトラフィックを搬送する為に使用されている割合である。例えば、“70%”の使用率制約が、あるパスグループに対して課されており、現在のモニタ結果が75 %を示しておれば、プライマリストレージシステム１００ａは本パスグループを使用する事は出来ない。

【００５２】実施例次第で、プライマリストレージシステム１００ａ内のプロセッサ、ハードウエア、及び/又は、ソフトウエア機構が各ポートの使用率を監視する。例えば、プロセッサが一定時間中、例えば１秒間にあるポートを通って転送される時間をモニタする。その後、各ポート対応の転送時間が加算され、全体の合計が当該パスグループの使用率になる。プライマリストレージシステム１００ａは、使用率制約が満足されればステータス４４０を“使用可能”に、使用率制約が満足されなければステータス４４０を“一時的使用不能”に設定する。

【００５３】プライマリストレージシステム１００ａは使用率監視を継続して、使用率が制約以下に落ちれば、ステータス４４０を“使用可能”に変更する。あるパスグループのステータス４４０が“一時的使用不能”を示し続けている間も、プライマリストレージシステム１００ａは当該パスグループをテスト使用する為に規則的間隔で選択する。プライマリストレージシステム１００ａはテスト目的以外では、他のパスグループを使用する。

【００５４】エラー率制約は、プライマリストレージシステム１００ａが１つのパスグループについて使用出来る最大エラー率を制限する。例えば、“10%”のエラー率制約が、あるパスグループに対して設定されており、モニタ結果、現在のエラー率が15%を示していれば、プライマリストレージシステム１００ａは、本パスグループを選択する事は出来ない。実施例によっては、プライマリストレージシステム１００ａ内のプロセッサ、ハードウエア、及び/又は、ソフトウエア機構が各ポートのエラー率を監視する。プロセッサが一定時間中、例えば１分間にあるポートを通って転送される転送数とエラー数をモニタする。その後、各ポート対応のこれらの数が加算され、全体の合計エラー数を全体の転送数で割った値が当該パスグループのエラー率になる。

【００５５】プライマリストレージシステム１００ａはエラー率を継続して監視し、エラー率が制約以下に落ちれば、ステータス４４０を“使用可能”に変更する。あるパスグループのステータス４４０が“一時的使用不能”を示し続けている間も、プライマリストレージシステム１００ａは当該パスグループをテスト使用する為に規則的に間隔をおき選択する。プライマリストレージシステム１００ａはテスト目的以外では、他のパスグループを使用する。

【００５６】外部制約は、プライマリストレージシステム１００ａに対して、当該システムの外部機構より提供されるパスグループに関する使用可能性情報に基づいて、パスグループの選択に制限を課す。例えば、ネットワーク１４０aを監視するネットワークモニタは管理コンソール１２０ａに接続され、本管理コンソール１２０ａを通して、プライマリストレージシステム１００ａの使用可能条件を設定する。本ネットワークモニタは、例えば、使用率、サービス不能のルーター数、エラー率、パケット喪失率、パケット衝突率等を監視する。もし、本ネットワークモニタが異常状態を発見したら、プライマリストレージシステム１００ａに対して報告し、ネットワーク１４０ａが使用可能状態に戻るまで、プライマリストレージシステム１００ａがステータス４４０を“使用不能”にセットする。

【００５７】他の外部制約の例としては、ユーザからの介入がある。たとえば、ユーザは定期メンテナンス等の為に、ネットワーク１４０aを一時的に使用不能にすることがある。本メンテナンス実行前に、ユーザはプライマリストレージシステム１００ａ上で、管理コンソール１２０ａを使用して、当該ネットワーク１４０ａのステータス４４０を“使用不能”にする。本メンテナンス終了後、ユーザはステータス４４０を再び“使用可能”に設定する。

【００５８】図６及び図７は、本発明の１実施例での代表的なユーザインターフェースのダイアグラムである。ユーザがネットワークトラフィックに制約を課す為には、プライマリストレージシステム１００ａに対して、制約情報を与えることができる必要性がある。図６に示される管理画面６００は、管理コンソールでのその為のユーザインターフェースである。ユーザが管理アイコンをクリックすると、管理コンソールで、管理画面６００が表示される。サーバボックス６１０がサーバとボリュ

ームの関係を表示する。図６の例では、”Juno”と称されるサーバが”/dev/rdsk/c1t1d0”、”/dev/rdsk/c1t2d0”と称される２つのボリュームを有する。ユーザがこれらのボリューム内の１つを選択すると、デバイス情報ボックス６２０が現れる。

【００５９】図６では、デバイス情報ボックス６２０はデバイス”/dev/rdsk/c1t2d0”の情報を表示している。デバイス情報ボックス６２０はストレージシステム情報６３０、デバイス情報６４０、リモートコピー情報６５０を表示する。リモートコピー情報６５０は当該ボリュームがミラーを形成しているか否かを、また、ミラーを形成しているならその状態を示す。図６でのPAIR状態は、プライマリ及びセカンダリボリュームがミラーを形成していることを示している。

【００６０】リモートストレージシステム情報は、ペア化されたストレージシステムの製造番号を示す製番６５２と当該ペアのロケーションを示すロケーション６５３を含む。ユーザが製番６５２に対応する３角マークをクリックすると、ストレージシステム情報６３０で記されるローカルストレージシステムに接続された使用可能なストレージシステムの情報が表示される。

【００６１】ポート情報は、ローカルストレージシステムに対して定義された全てのパスとその状態を表示するパスグループ６５４を含む。あるパスグループが選択されたリモートストレージシステムに接続していなければ、当該状態では“N/A”が示される。リモートストレージシステムに接続され、使用可能なら、状態“RDY”が示される。上から下へ向かっての表示順序はパスグループ使用の優先順位を表す。例えば図６において、パスグループ“T3 up to 5MB/s”は“Internet”より高いプライオリテイを持ち、プライマリストレージシステム１００ａはセカンダリストレージシステム１００ｂにデータ転送するときには、最初に“T3 up to 5MB/s”を選択する。ユーザは、本ユーザインターフェースを用いて、この優先順位を変更することが可能である。

【００６２】ユーザが、パスグループ６５４から１つのパスグループを選択すると、当該パスグループの情報がパスグループ名変更欄６５５、状態欄６５６、制約欄６５７に現れる。ユーザは、新しい名前をパスグループ名変更欄６５５に入力する事により、パスグループ名を変更できる。状態欄６５６は、選択されたパスグループの詳細状態を表示する。当該状態は、既に述べたように“使用可能”、“使用不能”、及び“一時的使用不能”のどれかとなる。ユーザが“使用不能”を選択すると、プライマリストレージシステム１００ａは、ボリューム内のデータ転送に本パスグループを使用しないので注意が必要である。

【００６３】本発明の実施例によっては、選択されたパスグループに対して、実に多くの種類の制約が制約欄６５７に現れる。ユーザが制約欄６５７に対応する三角ボタンをクリックすると、選択されたパスグループに対する制約が表示される。ある制約がそのパスグループに対して適用されているとき、制約名の左端にチェックマークが図6のように示される。ユーザが一つの制約を選択すると、(図６には示されない)対応画面が現れる。例えば、ユーザが図６上で“TP up to 5MB/s”制約を選択すると、図７で示される画面７００が現れる。例えば、画面７００の対話画面を使用して、スループット制約をセットするための必要情報を、ユーザは入力する事が出来る。

【００６４】ユーザが適用ボタン７５０をクリックすると、ユーザによってセットされた制約情報が読み込まれ、適用される。消去ボタン７６０をクリックすると、現在の制約情報は消去され、制約欄６５７のチェックマークは消える。ユーザは、図６の管理画面６００を使用して必要情報を入力する。ユーザが適用ボタン６６０をクリックすると、ユーザによって入力された制約情報が読み込まれ、適用される。本情報はプライマリストレージシステム１００ａによって適用され、ユーザによって入力された制約情報に従って、パス選択テーブル３００及びパスグループテーブル４００が生成、または、変更される。更に管理コンソール１２０ａは、ユーザによってパスグループ欄６５４に入力されたパスグループ名を、ポート番号セットにマップして、パスグループ名をポート番号に変換する。例えば、パスグループ“T3 up to 5MB/s”はポート０及び１に変換される。

【００６５】そして、管理コンソール１２０ａは、本ポート番号をボリューム番号と制約と共に、プライマリストレージシステム１００ａに送信する。

実装例

次に本発明について、多様な機能と特徴を有する使用例を用いながら説明する。このセクションは、これまで説明してきた制約を使用する多数の方法の幾つかを偏に説明目的に展開する。これらの事例では、高価なネットワークと安価なネットワークのタイプのように、異なる２種類のみのネットワークを使用する場合を例としているが、これは説明を簡単化するためであることにに注意をしてもらいたい。しかしながら、この分野の技術の専門家には明らかな如く、本発明が多様な実装に対応して、多様なタイプのネットワークを使用する多くの異なった構成が可能なことは明らかである。

【００６６】図８は本発明の１実施例で、一定値以下のスループット又は使用率で高価なネットワークを使用するときの代表的なプロセスを示すフローチャートである。図８に例示される実装においては、最大スループット又は使用率が一定値以下であれば高価なネットワークが使用される。ユーザが高価なネットワークに対して、スループット制約を指定すれば、高価なネットワークは指定されたスループット範囲内で使用される。更にまた、使用率制約を指定すれば、当該ネットワークは指定

使用率以下で使用される。本事例は、高価なネットワークを所定のデータスループット以下で使用するための代表的なものである。指定の最大スループットを越えた場合は、ユーザは追加の負担をしなければならないかもしれず、さもなくばネットワーク性能は大幅に低下する可能性がある。

【００６７】図8に示されるフローチャートは、プライマリストレージシステム１００ａが高価なネットワークを指定最大スループット又は使用率以下で使用して、スループット又は使用率が、制約での最大指定値を越えたら、安価なネットワークを使用するように、管理コンソールを使用して設定する為の制約戦略を示す。ステップ８００にて、図６、図７のユーザインターフェースを使用して、ユーザが高価なネットワークを事前設定された最大スループットまで使用する事にして、本高価なネットワークに第１のプライオリテイを与えることを指定する。次に、再び、図６、図７のユーザインターフェースを使用して、安価なネットワークを制約無しで使用する事にし、本ネットワークに第2のプライオリテイを与える。

【００６８】上記のステップにて、ユーザが制約戦略の設定を完了すると、プライマリストレージシステム１００ａは、図5のフローチャートに従って、セカンダリストレージシステム１００ｂに対してデータ転送を開始する。図5に関連して、既に述べられた様に、プライマリストレージシステム１００ａは、事前設定された最大スループットに達するまでは、高価なネットワークをデータ転送の為に選択する。一度、最大スループットに達すると、高価なネットワークはもはや制約を満たすことが出来ない為、プライマリストレージシステム１００ａは安価なネットワークを選択する。同様に、ユーザはステップ８００にて、使用率制約をセットする事も出来る。

【００６９】図９は、本発明の１実施例で夜間通信の安価なネットワークを使用する為の代表的なプロセスを示すフローチャートである。図9に示される実装例では、安価なパブリックネットワークが夜間操作の為に使用される。本実装例は、ユーザがパブリックネットワークを夜間に使用し、昼間帯では使用しない場合の代表的なものである。パブリックネットワークは、昼間は高トラッフィックになり、リモートコピーでの大量データ転送は、他のE-MailやWebアクセス等のサービスに影響が大きい為に、ユーザはパブリックネットワークの使用を夜間帯にのみ限定する。昼間帯でのパブリックネットワークの使用を回避する為に、ユーザは、時間制約をパブリックネットワークに設定する。

【００７０】例えば、プライマリストレージシステム１００ａがパブリックネットワークの使用を避けるために、ユーザは“9:00pmから6:00amの間のみ”の時間制約を設定する。図９のフローチャートは、ユーザが管理コンソール１２０ａを使用して、プライマリストレージシステム１００ａが、夜間通信の為に安価なネットワークを使用する為の制約戦略を示す。ステップ９００にて、ユーザは図６、図７のユーザインターフェースを用いて、安価なネットワークを夜間（例えば、9:00pmから6:00am）のみ使用出来るようにして、本ネットワークに第１のプライオリテイを与える。更に、ステップ９１０にて、再び図６，図７のユーザインターフェースを用い、高価なネットワークを制約無しに使用可能に設定し、これに第２のプライオリテイを与える。上記のステップにて、ユーザが制約戦略を設定した後は、プライマリストレージシステム１００ａは、図５のフローチャートに従って、セカンダリストレージシステム１００ｂにデータ転送を行う。

【００７１】ここでの記述と、図５からわかるように、プライマリストレージシステム１００ａは、夜間の9:00から朝の6:00の間は、安価なネットワークを選択する。プライマリストレージシステム１００ａは、他の時間帯は安価なネットワークは制約を満たさないため、高価なネットワークを選択する。同様に、ユーザはステップ９００にて、使用率制約を設定する事も出来る。

【００７２】図１０は、本発明の１実施例での他の代表的なシステム構成を示した図である。図１０で示される実施例では、プライマリストレージシステム１００ａは、非常時以外は安価なネットワークを使用する。本実施例は、高価なネットワークを使用時間ベースで支払い契約するケースの代表的なものである。本発明の多様な実施例では、多様な多くの非常ケースがあり得るであろう。以下非常ケースの代表的な事例を挙げる。

【００７３】(1) 安価なネットワークが高トラフィック状態で、プライマリストレージシステムが大量の未転送データを保有している場合。ネットワークを監視する外部ネットワークモニタが、安価なネットワークでの高トラフィック状態を検出したら、本ネットワークモニタはプライマリストレージシステム１００ａに通知する。すると、プライマリストレージシステム１００ａは、ネットワークモニタが安価なネットワークのトラフィックの減少を示すまでは、他のネットワークを使用する。

【００７４】(2) プライマリストレージシステムがあまりにも大量の未転送データを保有している場合。一般的には、安価なネットワークは高価なネットワークより低速である。従って、ネットワークに十分な帯域が無い限り、セカンダリストレージシステム１００ｂに転送すべきデータは、プライマリストレージシステム１００ａ中で蓄積される。安価なネットワークが低速のままであると、プライマリストレージシステムは、この事態を検出して、蓄積されたデータをセカンダリストレージシステムに転送する為に、高価ではあるがもっと高速なネットワークに切り替える。未転送データがたまり過ぎてプライマリストレージシステムのデータをセカンダリストレージシステムに対してミラーコピーが出来なくなってし

まう事態を避ける為に、プライマリストレージシステムは未転送データの蓄積量をモニタして、一定の限界値に達したら、高価ではあるがもっと高速なネットワークを使用する。

【００７５】(3)エラーが限界値を超えた場合。プライマリストレージシステム１００ａはネットワークを通したデータ転送でのエラー数を数え、エラー率を算出する。プライマリストレージシステム１００ａは、安価なネットワークでエラー数が限界値を超えたら、より高価なネットワークに切りかえる。限界値はユーザが決定することが出来る。プライマリストレージシステム１００ａは、高価なネットワークを使用中の間でも、テスト目的に安価なネットワークで定期的にデータ転送を行う。プライマリストレージシステム１００ａは、テスト目的以外では他のパスグループを選択する。プライマリストレージシステム１００ａは、安価なネットワークでのエラー数が限界値以下に低下すれば、高価なネットワークの使用を終了する。本技術は、安価なネットワークとして高負荷転送時に高いエラーが発生するTCP/IPプロトコルネットワークが使用されている場合には、特に有効である。

【００７６】(4) エラーが発生した場合。プライマリストレージシステムはネットワークを通した転送でのエラーの発生をモニタする。プライマリストレージシステムは、安価なネットワークでエラーが発生したら、代りに高価なネットワークに切りかえる。この技術は、プライマリストレージシステムが、まずは安価なネットワークを通じてデータ転送を試みる設定の実施例において有効である。プライマリストレージシステムは、安価なネットワークでうまく行かないときに、高価なネットワークに切りかえる。

【００７７】図１０では、２つのストレージシステム、即ち、プライマリストレージシステム１００ａとセカンダリストレージシステム１００ｂが、リモートストレージバックアップ用の１システムを構成する。ネットワークモニタ１０００は、ネットワーク１４０ａとネットワーク１４０ｂ及び管理コンソール１２０ａに接続される。ネットワークモニタ１０００は、ネットワーク１４０ａとネットワーク１４０ｂの活動を監視する。パス１０２０は、ネットワークモニタ１０００をネットワーク１４０ａとネットワーク１４０ｂに接続する。パス１０１０は、ネットワークモニタ１０００を管理コンソール１２０ａに接続する。パス１０２０とパス１０１０は、例えば、インターネットのような同じネットワークの一部でも良い。ネットワークモニタ１０００がネットワーク１４０ａで高トラッフィックを検出したら、管理コンソール１２０ａに通知する。

【００７８】図１１は、本発明の他の１実施例での代表的なネットワークモニタメッセージを示すダイアグラムである。図１１で示された代表的なネットワークモニタメッセージフォーマットで、ネットワーク名１１００は、図６のパスグループ６５４中に登録されたネットワーク名に対応する。例えば、“T3 up to 5MB/s”、“Internet”などが用いられる。ネットワーク名１１００の目的の１つは、管理コンソール１２０ａでネットワークが識別出来るようにする事である。警告１１１０は、ネットワークモニタ１０００がネットワーク監視中に検出した警告のタイプを示す。

【００７９】本発明の多様な実施例に対応して、多様で異なったタイプの警告が使用可能である。例えば、ある実施例では“過負荷”や“正常に復帰”などが報告される。“過負荷”警告は、ネットワークモニタ１０００が監視ネットワーク内で過負荷条件を検出したことを示す。“過負荷”警告には、ネットワークモニタ１０００がネットワーク監視中に測定した現在の使用率を含む。“正常に復帰”警告は、ネットワークモニタ１０００がネットワーク１４０ａが限界使用率以下のトラフィック量に復帰したことを示す。現日時欄１１２０は、ネットワークモニタ１０００が管理コンソール１２０ａにメッセージを発信した時刻を示す。

【００８０】図１２は、本発明の１実施例でのネットワークモニタを使用する為の代表的なプロセスを示すフローチャートである。図１２に示す実施例において、ネットワークモニタ１０００は、ステップ１２００から１２３０を実行し、管理コンソール１２０ａは、ステップ１２４０から１２６０を実行する。ステップ１２００にて、ネットワークモニタ１０００は、負荷変化、非常事態等の状態の変化を監視する。ステップ１２００にて、負荷変化が検出された場合、判定ステップ１２１０にて、正常から過負荷への変化か否かがチェックされる。状態変化が正常から過負荷への変化だったら、ステップ１２２０にて、ネットワークモニタ１０００は図１１で説明したフォーマットでのメッセージの警告欄１１１０に“過負荷”を書きこみ、本メッセージを管理コンソール１２０ａに送信する。逆に、ステップ１２１０にて、状態変化が過負荷から正常への変化だったら、ステップ１２３０にて、ネットワークモニタ１０００は図１１で説明したフォーマットでのメッセージの警告欄１１１０に“正常”を書きこみ、本メッセージを管理コンソール１２０ａに送信する。

【００８１】判定ステップ１２４０にて、ネットワークモニタ１０００から送られたメッセージを受信して、管理コンソール１２０ａは警告欄１１１０をチェックして、警告は“過負荷”か“正常に復帰”なのかを確認する。警告欄１１１０が“過負荷”を示しておれば、ステップ１２５０にて、管理コンソール１２０ａは、パスグループテーブル４００の当該ネットワークのステータス４４０に“一時的使用不能”をセットする。反対に、ステップ１２４０にて、警告欄１１１０が“正常”を示しておれば、ステップ１２６０にて、管理コンソール１２

０ａは、ステータス４４０に“使用可能”をセットする。図５のフローチャートで説明した通り、プライマリストレージシステム１００ａは、ステータス４４０が“一時的使用不能”か“使用不能”となっているネットワークを使用する事はない。

【００８２】従って、ネットワークモニタ１０００が過負荷を検出したネットワークはプライマリストレージシステム１００ａによって選択されることは無い。

【００８３】図１３は、本発明の１実施例で非常事態で高価なネットワークを使用する為の代表的なプロセスを示すフローチャートである。図１３の実施例では、稼動状態が平常な場合には安価なネットワークが使用される。ユーザが本安価なネットワークに対して、エラー率及び外部制約を課した場合、高価なネットワークは非常時のみに確保される。本実施例は、ユーザが非常時の場合のみに、高価なネットワークの使用を許可する代表的なものである。エラー制約条件を満たすことができない場合は、ネットワーク性能は大幅に低下して、セカンダリストレージシステムはプライマリストレージシステムのミラーイメージを維持出来なくなる事があり得る。

【００８４】ステップ１３００にて、ネットワークモニタ１０００は安価なネットワーク１４０ａを監視するよう設定される。図６，図７で述べられたユーザインターフェースを使用して、ネットワーク１４０ａの為の限界可動条件が設定される。一度、設定されると、ネットワークモニタ１０００は図１２で述べられた処理を実行する。

【００８５】ステップ１３１０にて、図６，図７で述べられたユーザインターフェースを使用して、ユーザは、安価なネットワーク１４０ａを事前設定された限界値によるエラー率制約及び外部制約により使用可能にして、本安価なネットワークに第１のプライオリテイを与える。外部制約は図１２で説明したように、ネットワークモニタ１０００よりのインプットをプライマリストレージシステム１００ａのパスグループテーブル４００に反映出来るようにする。ステップ１３２０にて、ユーザは高価なネットワーク１４０ｂを制約無しで使用可能にして、これに第２のプライオリテイを与える。この制約戦略により、プライマリストレージシステム１００ａは、ネットワークモニタ１０００が安価なネットワーク１４０ａで過負荷や異常状態を検出しない限り、本安価なネットワーク１４０ａを選択する。

【００８６】もし、ネットワークモニタ１０００が過負荷状態を検出したら、本情報は管理コンソール１２０ａに伝えられ、本状態は、パスグループテーブル４００内の安価なネットワーク１４０ａに対するステータス４４０に反映される。本ステータス欄４４０の変化は、プライマリストレージシステム１００ａをして、ネットワーク選択を替えさせ、安価なネットワーク１４０ａで過負荷状態が解消されるまで、高価なネットワーク１４０ｂを選択させる。これまで述べて来たことは、本発明の好適な実施例である。付記されているクレームで定義された本発明の範囲を逸脱する事なく、変更と修正が可能である事は言うまでもない。

【００８７】

【発明の効果】本発明により、従来の技術に勝る数々の便益が実現される。本発明は、プライマリとリモートストレージデバイスとの間の複数接続を通したデータ転送を管理する技術を提供する。顧客がネットワークアクセスを使用ベースで契約した場合、高価なネットワークは例外的状態の場合のみで使用されるため、費用を最小限にすることが出来る。

【図面の簡単な説明】

【図１Ａ】図１Ａは、本発明の１実施例での代表的なシステム構成を図示する

【図１Ｂ】図１Ｂは、本発明の１実施例での代表的なシステム構成を図示する。

【図２】図２は、本発明の１実施例でのパスとボリュームの代表的な関係を示す。

【図３】図３は、本発明の１実施例での代表的なパス選択テーブルを図示する。

【図４】図４は、本発明の１実施例での代表的なパスグループテーブルを図示する。

【図５】図５は、本発明の１実施例での代表的なパス選択プロセスを示すフローチャートである。

【図６】図６は、本発明の１実施例での代表的なユーザインターフェースのダイアグラムである。

【図７】図７は、本発明の１実施例での代表的なユーザインターフェースのダイアグラムである。

【図８】図８は、本発明の１実施例で一定以下のスループット、または使用率で高価なネットワークを使用する為の代表的なプロセスを示すフローチャートである。

【図９】図９は、本発明の１実施例で夜間通信で安価なネットワークを使用する為の代表的なプロセスを示すフローチャートである。

【図１０】図10は、本発明の一実施例での他の代表的なシステム構成を図示する.

【図１１】図１１は、本発明の他の１実施例での代表的なネットワークモニタメッセージを示すダイアグラムである。

【図１２】図１２は、本発明の１実施例でネットワークモニタを使用する為の代表的なプロセスを示すフローチャートである。

【図１３】図１３は、本発明の１実施例で非常事態で高価なネットワークを使用する為の代表的なプロセスを示すフローチャートである。

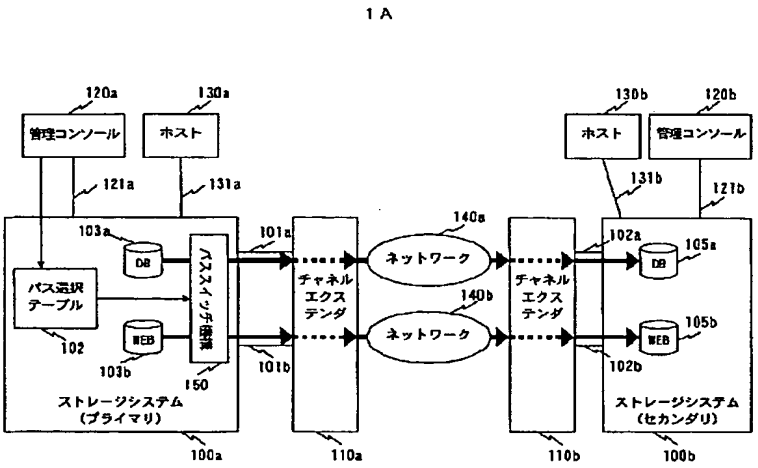
【符号の説明】

１００ａ･･･ストレージシステム(プライマリ)、１００ｂ･･･ストレージシステム(セカンダリ)、１０２･･･パス選択テーブル、１１０･･･チャネルエクステンダ、１２

０…管理コンソール、１３０…ホスト、１４０…ネットワーク、１５０…パススイッチ機構、１６０…ファイバインターフェース、１７０…IPインターフェース、１０００…ネットワークモニタ、１１００…ネットワーク名、インターネット、１１１０…警告、過負荷(使用率80%)

【図１Ａ】

【書類名】図面
【図１Ａ】



【図３】

【図３】

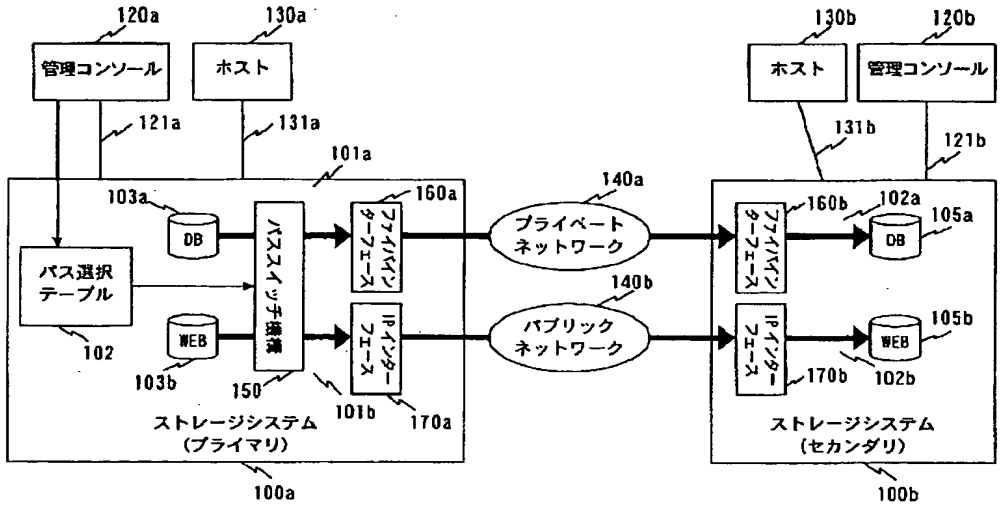
パス選択テーブル３００

310	320a	320b
ボリューム番号	パスグループ番号	パスグループ番号
0	0	NULL
1	1	0
2	0	NULL
3	1	NULL
4	1	0

【図１Ｂ】

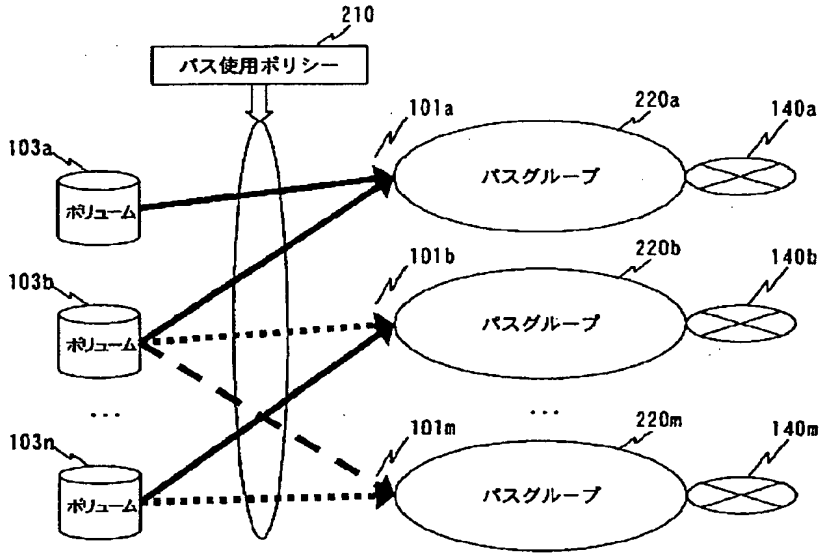
【図１Ｂ】

１Ｂ



【図２】

【図２】



【図４】

【図４】

パスグループテーブル４００

410 パスグループ番号	420 制　約	440 ステータス	430a リモートリンク番号	430b リモートリンク番号	430c リモートリンク番号
0	最大 5MB/s	使用不能	0	1	NULL
1	9:00pm から 6:00am の間のみ	使用可能	2	3	4
2	使用率 70%未満	使用可能	5	6	NULL
3	エラー率 5%未満	一時的使用不能	7	8	NULL
4	外部コントロール	一時的使用不能	9	10	11

【図１１】

【図１１】

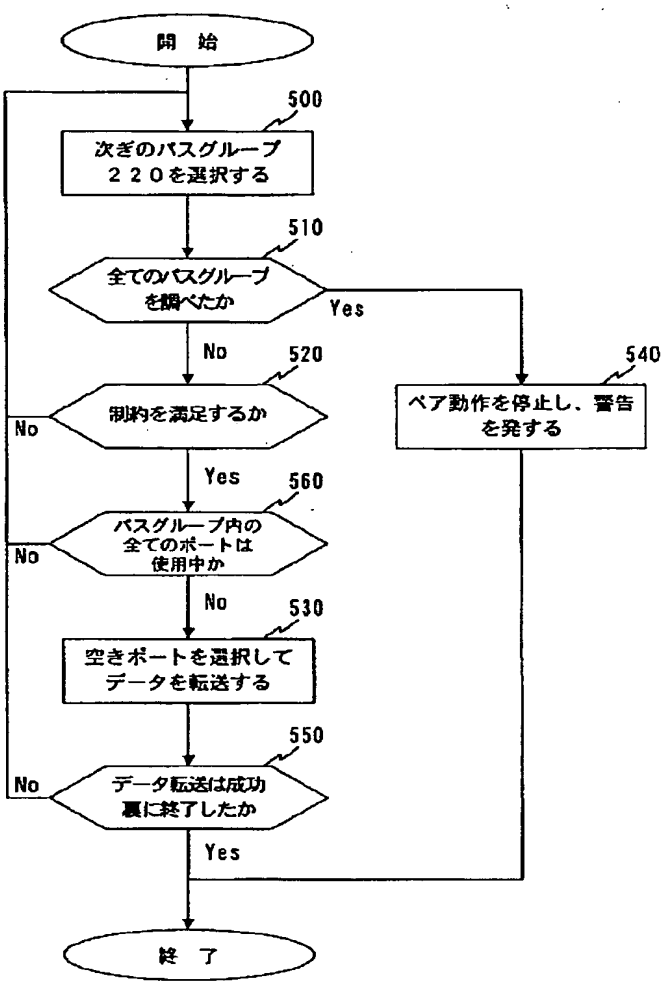
ネットワークモニタから管理コンソールへの情報送信

ネットワーク名	インターネット	1100
警　告	過負荷(使用率 80%)	1110
現 日 時	12-12-2000:13:46	1120

【図５】

【図５】

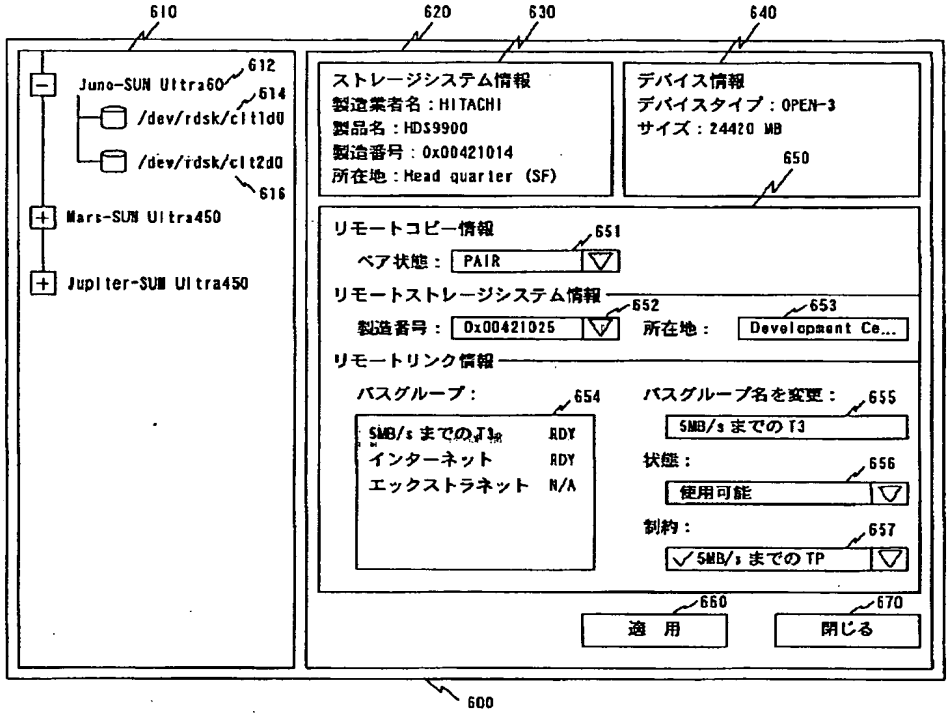
パス選択アルゴリズム



【図６】

【図６】

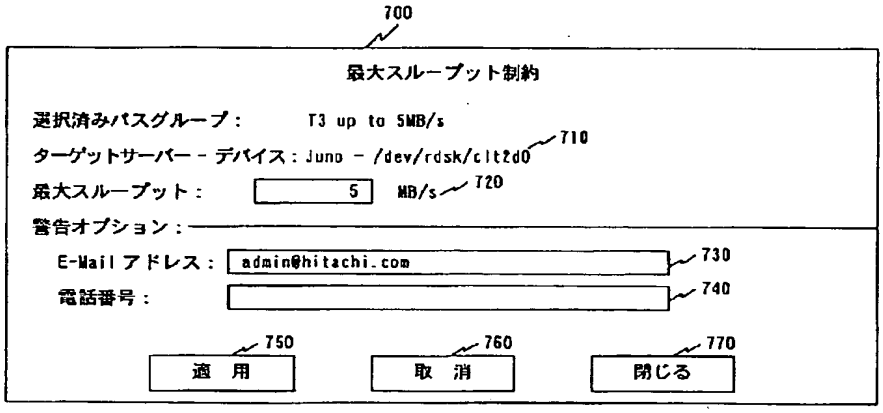
パス制約設定用ユーザインターフェース



【図７】

【図７】

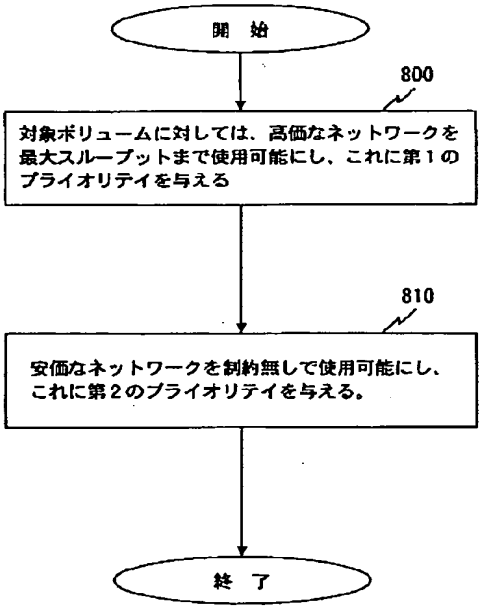
最大スループットの設定



【図８】

【図８】

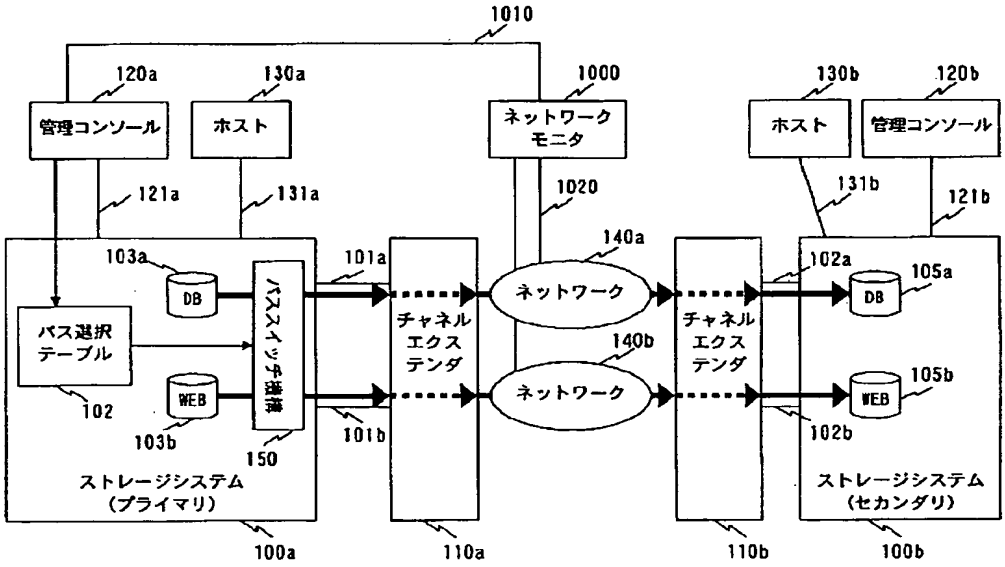
高価なネットワークを最大スループット未満、又は最大使用率未満で使用するためのフローチャート



【図１０】

【図１０】

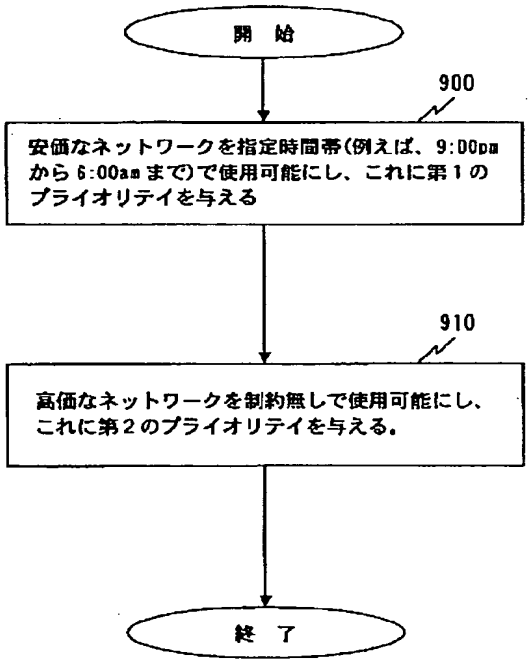
ネットワークモニタ付きシステム構成



【図９】

【図９】

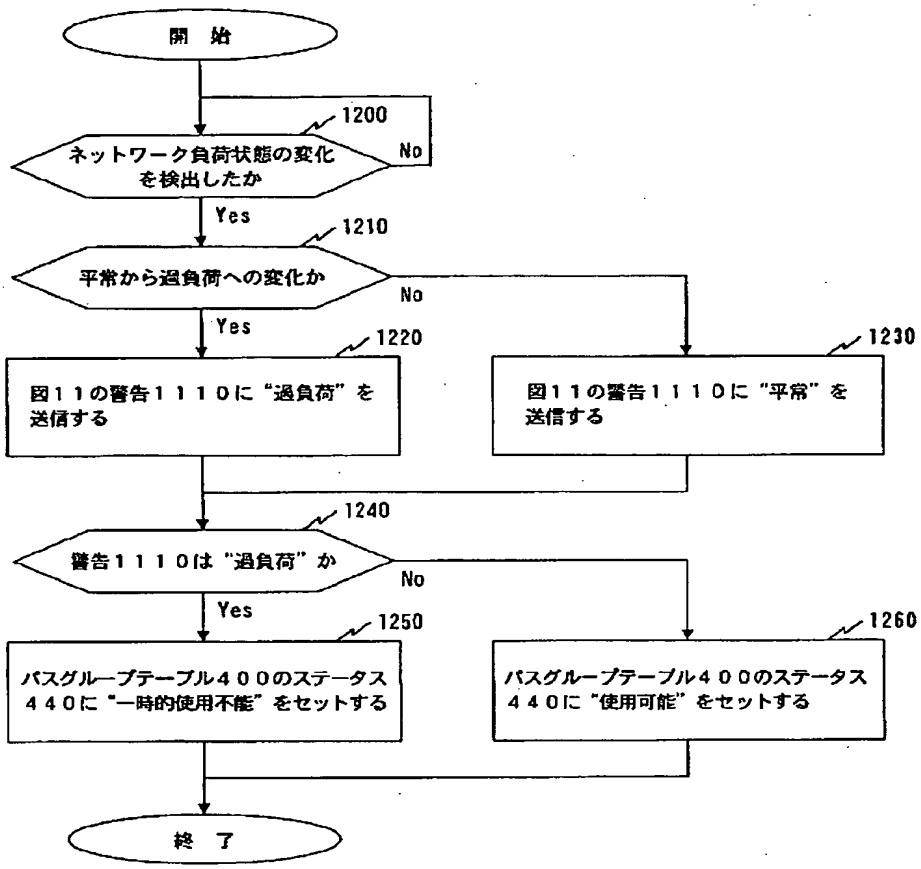
安価なネットワークを夜間のみに使用するためのフローチャート



【図１２】

【図１２】

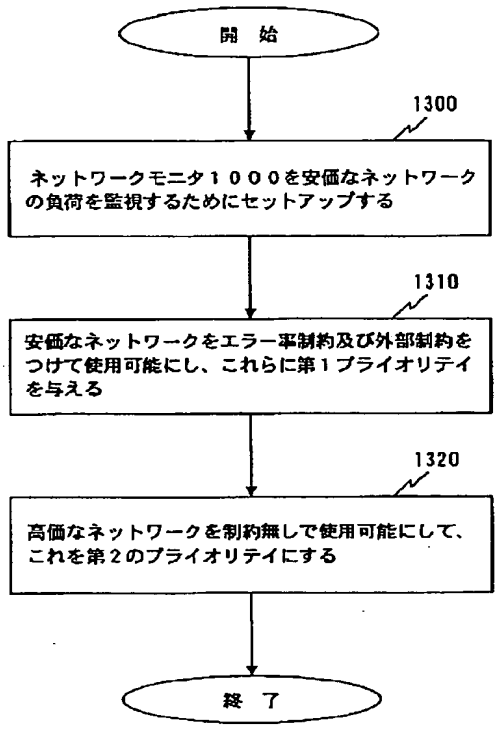
ネットワーク監視のフローチャート



【図１３】

【図１３】

高価なネットワークを非常時のみに使用するためのフローチャート



【外国語明細書】

Path Selection Methods for Storage Based Remote Copy

BACKGROUND OF THE INVENTION

The present invention relates generally to remotable data storage systems, and in particular to techniques for managing data flow over a plurality of connections between primary and remote storage devices.

The information technology revolution brings with it an ever increasing need for more storage capacity for business enterprises. It is expected that the average Fortune 1000 company's storage requirement will more than double in the coming years. In addition, growth has brought shortages of skilled persons in the information technology field. These challenges confront many companies facing the need to expand and improve their information technology assets. Increasingly, companies are turning to storage based remote copy as a method of coping with the need to prevent data loss from disaster. Remote copy creates and manages mirror images of storage volumes between a local, or primary storage system, and a remotable, or secondary storage system. The primary and secondary storage systems may be located at a far distance from one another. The two disk storage systems are connected by a network, through which updates on a local disk system are copied to the remote disk system. Nowadays, there are many types of networks that can connect the two storage systems performing remote copying. For example, one type of network can be a fast, reliable, secure and relatively more expensive network, such as, for example, a T3 private network. Another type of network is relatively more slow, insecure, and cheap, such as the Internet, for example.

Business critical applications, like on line transaction processing (OLTP) for banking, finance, flight reservation systems, and so forth, requires remote copy capabilities with low response times, high security, and high reliability. Other types of applications, like WEB mirroring, data warehousing, data center consolidation, bulk data transfer, and the like, do not have such requirements, because these applications generally do not need to copy data in real time.

While certain advantages to present remote copy technologies are perceived, opportunities for further improvement exist. For example, according to conventional remote copy technology, the network carrier companies charge customers based upon a required throughput, and sometimes offer pay per services for private networks. For example, a network carrier company may charge customers according to network bandwidth used per month. However, some remote copy users would like to reduce the costs associated with data connections and will be willing to accept operational limitations to do so. For example, users can lower expenses by using different networks for remote copy depending on application characteristics. A user could employ the Internet for web mirroring applications, but use a T3 network for OLTP for banking, for example. Users who would like to use the storage for backup purposes do not need a full-time data connection. However, conventional technology does not provide the capability to manage access to the data transmission services of a network carrier based upon the charges for the access. Further, issues such as security are important concerns to both the user and the network carrier. For the user, this means that valuable information assets can be protected by restricting access to the data being sent to remote storage. For the network carrier, this means that data integrity is preserved for each of its customers, and that no user receives access that is not authorized by the network carrier.

What is needed are improved techniques for managing data flow over a plurality of connections between primary and remote storage devices.

SUMMARY OF THE INVENTION

The present invention provides techniques for managing data flow over a plurality of connections between primary and remote storage devices. In a representative example embodiment, when the primary storage system copies data to the secondary storage system, it chooses one of a plurality of networks connecting it to the secondary storage system, depending upon a users' policy. Since networks have different characteristics, in terms of, for example, performance, security, reliability, and costs, the user can specify which network(s) are used under various circumstances, i.e., daytime operation, nighttime operation, normal operation, emergency, and so forth. The storage systems comprise a mapping of volumes and ports. When performing copy operations, the primary storage system finds a volume storing the data, and available ports by accessing the mapping. The mappings are based upon policies that are input by a user.

In a representative specific embodiment, the primary storage system can be configured to limit data transfers using a particular network to within a set maximum throughput. For example, if a user configures a 5MB/s of the maximum throughput for a network, the storage system uses the network only up to the threshold. When the 5MB/s threshold is reached, the primary storage system chooses ports connecting to other networks. This mechanism provides substantially improved performance when networks susceptible to overload are used for storage operations. In other specific embodiments, if network access is purchased on a pay per use basis, the user can limit expenses for using the pay per use network according to a budget, by limiting the use of the pay per use network to a particular throughput, say 5MB/s, in order to avoid incurring additional charges. In still further specific embodiments, the primary storage system may be configured to select ports connecting to inexpensive networks, except, for example, during daytime, when public networks experience high traffic volume. Further, the primary storage system that transfers remote copy data through a specific network may affect the performance of other network service, sometimes causing adverse conditions to corporate operations relying on these networks. Accordingly, the primary storage system can be configured to select from other, more expensive, i.e. private networks, for example, during the day to avoid these types of consequences.

In another representative specific embodiment, when a primary, i.e., inexpensive, network experiences a high traffic volume, an external network monitor that monitors traffic volume over the networks notifies the primary storage system. Then the primary storage system switches to other networks until the monitor informs the primary storage system that the primary network has returned to a low traffic volume. Another specific embodiment determines when the primary storage system has too much data pending transfer to the remote storage system. Generally, an inexpensive primary network is slower than an expensive secondary network. Accordingly, data to be transferred to the secondary storage system is accumulated in the primary storage system if traffic throughput of the primary network is insufficient to keep up with the data transfer demand of the primary storage system. If left unchecked, the secondary storage system will eventually be unable to maintain a mirror image copy of the primary storage system. To avoid this condition, the primary storage system monitors the quantity of data pending transfer that accumulates, and switches to a secondary, i.e., more expensive, network when the accumulated data exceeds a threshold.

In a further representative specific embodiment, a method for minimizing cost of network access by a storage apparatus is provided. The method comprises specifying a first network to be used for transferring data. Specifying a constraint for the first network is also part of the method. In various specific embodiments, the constraint comprises at least one of a throughput, a busy rate, an error rate, and a presence of an error, for example. However, other types of constraints are also used in various specific embodiments. The method also includes specifying a second network to be used for transferring data. Transferring data using the first network when conditions in the first network are in accordance with the constraint, otherwise transferring data using the second network is also included in the method. In a specific embodiment, the method further comprises transferring a portion of the data using the first network even when conditions in the first network are not in accordance with the constraint as a test, monitoring conditions in the first network during the test; and returning to transferring data using the first network when the test reveals that conditions in the first network are again in accordance with the constraint. In specific embodiments, the first network may be relatively less expensive to use than the second network, and/or the first network is a public network and the second network is a private network. When the user specifies the networks and constraints, the use can make the first network a higher priority network than the second network, or configure the apparatus such that detecting an abnormal condition in the first network and thereupon transferring data using the second network, for example.

Another strategy monitors an error count, such as a percentage error rate. The primary storage system monitors how many errors occur during data transfer through the network, and calculate an error rate. When the error rate becomes too great, which can be determined by exceeding a threshold, for example, the primary storage system switches to an expensive network. The threshold error rate can be determined from a customer's policy, for example. While using the expensive network, the primary storage system will also attempt to use the inexpensive network in order to continue to monitor the status of the inexpensive network. The primary storage system will discontinue using the expensive network if the error rate for the inexpensive network falls below the threshold. When TCP/IP protocol is used as the inexpensive network, a high occurrence of errors often indicates a high traffic volume in the network.

A still further strategy switches to expensive networks as an alternate data path to the inexpensive networks when an emergency occurs. According to this strategy, the primary storage system transfers data using the inexpensive network. But, if this fails, the primary storage system switches to the more expensive network.

In another representative embodiment, a method for selecting a network is provided. The method comprises monitoring one or more conditions in a plurality of networks. Comparing the one or more conditions in the plurality of networks to one or more user provided policies; and selecting one or more ports connecting to the plurality of networks are part of the method. In a specific embodiment, the monitoring one or more of conditions in the plurality of networks comprises using a network monitor to detect a condition within at least one of the plurality of networks, and thereupon set a value in a status indication, and the selecting of one or more of ports connecting to the plurality of networks comprises determining based upon a status indication whether to select a port from the one or more of ports connecting the plurality of networks. Each of the plurality of networks has one or more of user provided policies associated with it. In one specific embodiment, the method also comprises associating the plurality of networks with a plurality of path groups and then associating the one or more policies based upon the one or more path groups.

In a still further representative specific embodiment, a storage apparatus is provided. The storage apparatus comprises one or more disk drives; a memory that is operable to contain path selection information; a plurality of ports that provide switchable connection to a plurality of networks; and a processor. Each of the plurality of networks has one or more user provided policies associated with it. Representative policies include, for example, a threshold, a maximum, a minimum, an average, a mean, a limit, a constraint, a priority, and a target. The processor, based upon monitoring of one or more conditions in the plurality of networks, selects at least one of the ports connecting the plurality of networks, based upon a comparison of the conditions in the plurality of networks to the plurality of user provided policies. Representative conditions include, for example, a throughput, a busy rate, an error rate, and a presence of an error. In specific embodiments, the storage apparatus further comprises a plurality of status indications, each of which is associated with one of the networks. The processor determines based upon the status indication whether to select a port from the one or more ports connecting to the plurality of networks. Representative statuses include, for example, available,

temporarily unavailable, and unavailable. In a specific embodiment, a network monitor is also provided, which is operable to detect a condition within one or more networks, and thereupon to set a value in the status indication. Further, in some specific embodiments, the networks are grouped into a plurality of path groups, so that policies may be associated with the networks in a particular path group. Further, the disk drives may be divided into volumes, and the each of the volumes is permitted to access networks of one or more of the path groups.

Numerous benefits are achieved by way of the present invention over conventional techniques. Specific embodiments according to the present invention provide techniques for managing data flow over a plurality of connections between primary and remote storage devices. If a customer purchases network access on a pay per use basis, these techniques keep expenses lower, since expensive networks are used during exceptional conditions. While the present invention has been described with reference to specific embodiments having a first and a second network, this is intended to be merely illustrative and not limiting of the wide variety of specific embodiments provided by the present invention.

These and other benefits are described throughout the present specification. A further understanding of the nature and advantages of the invention herein may be realized by reference to the remaining portions of the specification and the attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Figs. 1A-1B illustrate drawings of representative system configurations in a specific embodiment of the present invention.

Fig. 2 illustrates a drawing of a representative relationships between paths and volumes in a specific embodiment of the present invention.

Fig. 3 illustrates a drawing of a representative path selection table in a specific embodiment of the present invention.

Fig. 4 illustrates a drawing of a representative path group table in a specific embodiment of the present invention.

Fig. 5 illustrates a flowchart of a representative path selection process in a specific embodiment of the present invention.

Fig. 6 illustrates a diagram of a representative user interface in a specific embodiment of the present invention.

Fig. 7 illustrates a diagram of a representative user interface in a specific embodiment of the present invention.

Fig. 8 illustrates a flowchart of representative processing in an implementation that uses an expensive network below a particular throughput or busy rate in a specific embodiment of the present invention.

Fig. 9 illustrates a flowchart of representative processing in an implementation that uses an inexpensive network during night operations in a specific embodiment of the present invention.

Fig. 10 illustrates a drawing of another representative system configuration in a specific embodiment of the present invention.

Fig. 11 illustrates a diagram of a representative network monitor message in another specific embodiment of the present invention.

Fig. 12 illustrates a flowchart of representative processing in an implementation that uses a network monitor in a specific embodiment of the present invention.

Fig. 13 illustrates a flowchart of representative processing in an implementation that uses an expensive network in emergency situations in a specific embodiment of the present invention.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

The present invention provides improved techniques for managing data flow over a plurality of connections between primary and remote storage devices.

Remote copy technology provides mirror image copies of one of a pair of disk systems to the other member of the pair. The two disk systems are interconnected by ports and located at some distance from one another. The remote copy system keeps a mirror image of disks located in the local, or primary system. The mirror image is stored in a remote, or secondary disk system. The local disk system copies data on a local disk of the pair. When a host updates data on the local system's disk, the local disk system transfers a copy of the data to the remote system through a series of ports and network links. Accordingly, no host operation is required to maintain a mirror image of a volume in the local system. For further description of representative remote copy systems in the

art, reference may be had to a variety of references, such as U.S. Patent Nos. 5,459,857 and 5,544,347.

Various types of methods exist for transferring data between the local and remote disk systems. In one type, called a "synchronous mode," the local disk system transfers data to the remote disk system before indicating that a write request for the data from a host is complete. In another type, called a "semi-sync mode," the local disk system indicates that the write request for data from a host is complete and then transfers the write data to the remote disk system. In both of these types of modes, succeeding write requests from the host are not processed until a previous data transfer is indicated to the host as finished. In an "adaptive copy mode" by contrast, data which is pending copy to the remote disk system is stored in a memory in the primary disk system, and transferred to the remote disk system when the local disk system and/or ports are available for the copy task. Accordingly, disk write operations by the host system to the primary system can continue without pause for completion of the copy operation to the remote storage system. For further description of representative transfer modes in remote copy systems in the art, reference may be had to a variety of references, such as U.S. Patent No. 5,933,653.

Figs. 1A-1B illustrate drawings of representative system configurations in a specific embodiment of the present invention. As shown in Fig. 1A, two storage systems, which are named a primary storage system 100a and a secondary storage system 100b, comprise one configuration for using a remote storage backup system. Each of the primary storage system 100a and the secondary storage system 100b comprise one or more volumes that store data. The storage systems 100a and 100b have processors which execute programs, and a memory for storing control data and tables for the programs. During operation, data stored on volumes of the primary storage system 100a is copied to identical volumes in the secondary storage system 100b. This operation is sometimes referred to as "mirroring" or "mirror imaging." For example, information stored on volumes 103a and 103b of the primary storage system 100a may be mirrored on the volumes 105a and 105b of the secondary storage system 100b. The primary storage system 100a and the secondary storage system 100b may be under the control of a single entity, or alternatively, a service provider may own a storage system which is used to provide backup services to the owner of the primary storage system 100a. Additionally, in some embodiments, the role of primary copy and secondary, or backup copy, may be

reversed or even shared between the two storage systems. In these embodiments, the secondary storage system 100b may mirror some of the volumes of primary storage system 100a, and the primary storage system 100a may mirror some of the volumes of the secondary storage system 100b.

One or more host systems, such as host 130a and host 130b, connect to at least one of the primary storage system 100a and the secondary storage system 100b by a channel path 131a, and a channel path 131b, respectively. In example specific embodiments, channel paths 131a and 131b are implemented using SCSI, Fibre Channel, ESCON, and the like. The host systems 130a and 130b access data stored on the volumes 103a and 103b in the primary storage system 100a and the secondary storage system 100b, respectively, through channel paths 131a and 131b, respectively.

Management consoles 120a and 120b connect to the primary storage system 100a and the secondary storage system 100b, respectively, by paths 121a and 121b, respectively. In an example embodiment, the paths 121a and 121b may be LAN, proprietary path, SCSI, Fibre Channel, ESCON, and the like. An administrator inputs policies for creating path selection table 102 through management console 120a.

In a representative specific embodiment, the network 140a is a public, low performance, low security, network that is relatively low in cost to use. In an example embodiment, network 140a is the Internet. As used herein, the term "public" is used to refer to networks that are accessible by virtually anyone who is certified (or sometimes uncertified). The network 140b is a private, high performance, high security network that is relatively more expensive to use. In an example embodiment, the network 140b is a T3 communication line. As used herein, the term "private" is used to refer to networks that are dedicated to a particular user, or group of users, and that others cannot access. The term "public" is used to refer to all other networks.

This present invention is described using simplified representative embodiments, in which just two types of networks provide connections between the primary storage system 100a and secondary storage system 100b, for clarity. However, these simplified examples are intended to be merely illustrative for the purposes of explanation, rather than limiting of the present invention. In many specific embodiments, three or more different types of networks are used in a manner similar to that described herein with reference to these specific embodiments.

In Fig. 1A, a plurality of channel extenders 110a and 110b provide protocol conversion between ports, such as ports 101a and 101b, and the networks 140a and 140b. For example, if a port 101a is a SCSI type interface, and network 140a is the Internet, then the channel extenders 110a and 110b convert data from the SCSI format to the TCP/IP protocol, and vise versa. One or more ports, such as ports 101a and 101b, connect the primary storage system 100a and the channel extender 110a. The channel extender 110a provides connection to the networks 140a and 140b. Port 101a provides connection to network 140a, while port 101b provides connection to network 140b through the channel extender 110a. One or more ports 102a and 102b also connect the secondary storage system 100b and the channel extender 110b. The channel extender 110b provides connection to the networks 140a and 140b.

Fig. 1B illustrates an alternative specific embodiment, in which the primary storage system 100a and/or secondary storage system 100b support various types of protocols. Accordingly, the respective channel extenders 110a and 110b are not required. In this specific embodiment, the ports 101a and 101b in the primary storage system 100a connect directly to the networks 140a and 140b using one or more interfaces, such as a fibre interface 160a and an IP interface 170a, for example. Analogously, ports 102a and 102b in the secondary storage system 100b connect directly to networks 140a and 140b via fibre interface 160b and IP interface 170b, respectively.

Fig. 2 illustrates a drawing of a representative relationships between paths and volumes in a specific embodiment of the present invention. As shown in Fig. 2, volumes 103a, 103b, ..., 103n within the primary storage system 100a are capable of sending data via one or more path groups, such as path groups 220a, 220b, ..., 220m. Each path group comprises one or more ports, such as ports 101a and 101b, that connect to a network, such as network 140a, for example. In Fig. 2, path group 220a comprises port 101a, connecting to network 140a, path group 220b comprises port 101b, connecting to network 140b, and so forth.

A path using policy 210 maps a volume and one or more path groups, and defines priority for using paths when transferring data to the secondary storage system 100b. For example in Fig. 2, when the primary storage system 100a transfer data on volume 103b to the secondary storage system 100b, it selects a port 101b in path group 220a. However, if the path group 220a is not available for some reason, then it selects path group 220b. If the path group 220b is not available, then it selects path group 220m.

For another example in Fig. 2, a volume 103a is allowed to use paths in only path group 220a. In a specific embodiment, the path using policy 210 is implemented using tables, as illustrated in Fig. 3 and Fig. 4, which are described herein.

Fig. 3 illustrates a drawing of a representative path selection table in a specific embodiment of the present invention. As shown in Fig. 3, a path selection table 300 maps a volume number 310, which corresponds to a volume, such as volumes 103a and 103b in Fig. 1, and one or more path group numbers 320a and 320b, which correspond to path groups, such as path groups 220a and 220b in Fig. 2, for example. The volume number 310 is unique to each volume within a storage system. For example, the primary storage system 100a, which comprises volumes 103a and 103b, will have unique volume numbers 310 corresponding to the volumes 103a and 103b in the path selection table 300.

The path group numbers 320a and 320b are unique to each path group defined for a storage system. When two or more entries for the path group number exist, such as 320a and 320b, for example, the number of the entries corresponds to the number of different networks connected to the storage system. The preceding path group numbers, 320a, have a higher priority than succeeding path group numbers, 320b. When transferring data to the secondary storage system 100b, the primary storage system 100a selects a path group having a higher priority. For example, in Fig. 3, the path group number 320a has higher priority than path group number 320b. In the second row of table 300, when transferring data on volume with number 1, primary storage system 100a selects a port in path group with number 1 rather than path group with number 0.

If there are fewer path groups than entries for path groups in path selection table 300, then a NULL string is stored in the remaining entries in the path selection table. For example in Fig. 3, the volume number 0 is allowed to use path group number 0, but no other path groups. So, a NULL is stored in the path group number 320b for the volume 0.

Fig. 4 illustrates a drawing of a representative path group table in a specific embodiment of the present invention. As shown in Fig. 4, a path group table 400 provides information about the path using policy 210, and maps path groups and ports. The path group table 400 comprises a path group number 410, which is a unique number assigned to each of the path groups for a particular storage system. A constraint 420 holds constraints that apply to the use of the paths within the path groups. For example,

the constraint 420 stores "Max 5MB/s" for path group 0. Accordingly, the paths in the path group 0 must not exceed 5MB/s for transferring data. One or more constraints can be registered to a particular path group. A variety of types of constraints can be used in various specific embodiments according to the present invention. Representative examples of specific constraints are described herein below. A port number 430a, 430b, and 430c each hold a number corresponding to a port in the path group. Each port has a unique port number within a storage system 100.

If there are fewer ports in a path group than there are entries in the path group table 400, then a NULL string is stored into the vacant entries. For example, the path group number 0 has only two ports, port number 0 and port number 1. Therefore, a NULL is stored in the port number 430c.

A status 440 holds a current status of a path group. In a specific embodiment, the status takes values such as "available," "unavailable," or "temporarily unavailable." The status of "available," indicates that the primary storage system 100a can use the corresponding port in the path group. The status of "unavailable" indicates that the primary storage system 100a cannot use the corresponding port. The status of "temporarily unavailable" allows the primary storage system 100a to attempt to use a path group 220 for a certain interval, e.g. once per minute, in order to check availability. For example, if constraint 420 comprises "Error rate less than 5%" and status 440 shows "temporarily unavailable," then data is transferred via the path group 220a once per minute, for example. The primary storage system 100a monitors the results. When the error rate falls below 5%, for example, the primary storage system 100a changes the status 440 to "available."

Fig. 5 illustrates a flowchart of a representative path selection process in a specific embodiment of the present invention. As shown in Fig. 5, when a request to transfer data to the secondary storage system 100b arises, the primary storage system 100a executes a plurality of steps. In a step 500, the primary storage system 100a selects a path group to transfer data to the secondary storage system 100b, by accessing path selection table 300. Since primary storage system 100a knows the volume storing data to be transferred, it determines a row corresponding the volume in the path selection table 300. Then, it selects path group number 320a in the first iteration. If the path group 220a does not satisfy constraints in a step 520, then the primary storage system 100a selects path group number 320b in the second iteration.

In a step 520, the path group is examined to determine whether all constraints are satisfied. If all path groups listed in the path selection table 300 do not satisfy the constraints, then processing proceeds with a step 540. In step 540, the primary storage system 100a suspends the mirroring operations between the pair of volumes in the primary storage system 100a and the secondary storage system 100b, and reports a warning to a user.

If there is a path group that satisfies the constraints in step 520, then, at a step 560, a check whether all ports in the path group are busy is performed. A port is busy when the primary storage system 100a is transferring data using the port. If there's a port that is idle, then, in a step 530, the primary storage system 100a selects the idle port, and transfers data through the port. Next, in a step 550, a check is performed to determine if the data transfer successfully completed. If the data transfer completed successfully, then processing is finished. Otherwise, control proceeds back to step 500, in which the primary storage system 100a tries another path group.

Constraints

A variety of types of constraints may be used in specific embodiments of the present invention. The following are representative examples of constraints that may be used in various specific embodiments. This list is not intended to be exhaustive, but rather, illustrative of some of the many different types of constraints that are used in various specific embodiments of the present invention.

A time constraint limits the time when the primary storage system 100a is allowed to use a particular path group. For example, if a time constraint of "9:00 pm to 6:00 am only" is active for a particular path group, and the current time is 8:00 am, then the primary storage system 100a must not use paths in the particular path group. The primary storage system 100a comprises a time clock, which is used to determine if the time is within the bounds of a time constraint, if a time constraint exists for a particular path group. The primary storage system 100a checks the time clock on a regular basis (e.g. once per minute). When the time constraint is satisfied, the primary storage system 100a changes the status 440 to "available." Similarly, when the time constraint is no longer satisfied, then the primary storage system 100a changes the status 440 to "unavailable."

A throughput constraint limits the maximum throughput that the primary storage system 100a is allowed to use from a particular path group. For example, if a throughput constraint of "5MB/s" has been set for a particular path group, and the current result of monitoring shows a throughput of 5.3MB/s being used, then the primary storage system 100a must not use paths in the particular path group. In various specific embodiments, processors, hardware, and/or software mechanisms within the primary storage system 100a monitor throughput of each port. In a specific embodiment, processors monitor the quantity of data transferred by a particular port during a specific time interval, such as every second. Then, a sum of the quantities monitored by each processor is computed. This sum indicated the throughput for the particular path group comprising the ports. When the throughput constraint is satisfied, the primary storage system 100a changes the status 440 to "available." Similarly, when the throughput constraint is no longer satisfied, then the primary storage system 100a changes the status 440 to "temporarily unavailable."

The primary storage system 100a continues to monitor throughput, and will set the status 440 to "available" when the throughput falls below the constraint. In a specific embodiment, while the status 440 continues to show that a particular path group is "temporarily unavailable," the primary storage system 100a selects the particular path group at regular intervals, to perform a trial data transfer. The primary storage system 100a selects the remaining path groups to perform non-trial data transfers.

A busy rate constraint limits the maximum "busy rate" that primary storage system 100a is allowed to use a particular path group. As used herein, the term "busy rate" refers to a percentage of total capacity of a network line which is being used to carry traffic. For example, if a busy rate constraint of "70%" has been set, and the current monitoring results indicate that a particular path group is 75% busy, then the primary storage system 100a must not select new paths in that particular path group. In various specific embodiments, processors, hardware, and/or software mechanisms within the primary storage system 100a monitor throughput of each port. In a specific embodiment, processors monitor the time that each port is used to transfer data during a specific interval, such as every second. Then, a sum of the time determined by monitoring each port is computed. This sum indicates the busy rate for the particular path group comprising the ports. When the busy rate constraint is satisfied, the primary storage system 100a sets the status 440 to "available." Similarly, when the busy rate

constraint is no longer satisfied, then the primary storage system 100a changes the status 440 to "temporarily unavailable."

The primary storage system 100a continues to monitor busy rate, and will set the status 440 to "available" when the busy rate falls below the constraint. In a specific embodiment, while the status 440 continues to show that a particular path group is "temporarily unavailable," the primary storage system 100a selects the particular path group at regular intervals, to perform a trial data transfer. The primary storage system 100a selects the remaining path groups to perform non-trial data transfers.

An error rate constraint limits the maximum error rate that the primary storage system 100a is allowed to use a particular path group. For example, if the error rate constraint of "10%" has been set, and the current results of monitoring indicate that an error rate of 15% is present in a particular path group, then the primary storage system 100a must not select new paths in that particular path group. In various specific embodiments, processors, hardware, and/or software mechanisms within the primary storage system 100a monitor error rate of each port. For example, processors count the total number of transfers and the total number of errors for a port during a specific time interval, such as every minute. Then the sum of these results for each port is computed. The sum indicates the total number of transfers and errors. Dividing the total errors by the total transfers shows the error rate.

The primary storage system 100a continues to monitor the error rate, and will set the status 440 to "available" when the error rate falls below the constraint. In a specific embodiment, while the status 440 continues to show that a particular path group is "temporarily unavailable," the primary storage system 100a selects the particular path group at regular intervals, to perform a trial data transfer. The primary storage system 100a selects the remaining path groups to perform non-trial data transfers.

An outboard constraint limits the selection of paths by the primary storage system 100a based upon information about the availability of path groups provided by mechanisms outside of the primary storage system 100a. For example, a network monitor that monitors network 140a, is connected to the management console 120a, and sets the availability of the primary storage system 100a via management console 120a. The network monitor monitors, for example, a busy rate, a number of routers that are out of service, an error rate, a rate of packet loss, a collision rate of packets, and the like. If the network monitor finds abnormal conditions, then it informs the primary storage system

100a, which sets the status 440 to "unavailable" until the network 140a becomes available.

Another example of an outboard constraint is intervention by a user. For example, users may temporarily make network 140a unavailable to perform routine maintenance, and the like, for example. Before performing maintenance, the user sets the status 440 to "unavailable" for the network 140a in the primary storage system 100a using the management console 120a. After completing the maintenance, the user sets the status 440 to "available" once again.

Figs. 6 and 7 illustrate diagrams of a representative user interface in a specific embodiment of the present invention. In order for users to apply constraints to networks traffic, users need to be able to provide constraint information to the primary storage system 100a. As shown in Fig. 6, a management window 600 provides a user interface to a user at the management console. When a user clicks a management icon, the management window 600 is displayed on the management console to the user. A server box 610 shows the relationship between servers and volumes. In the example of Fig. 6, the server named "Juno" has two volumes named "/dev/rdsk/c1t1d0" and "/dev/rdsk/c1t2d0." If a user selects one of these volumes, then the device information box 620 appears. As shown in Fig. 6, the device information box 620 shows the information for the device "/dev/rdsk/c1t2d0." The device information box 620 provides storage system information 630, device information 640, and remote copy information 650. In the remote copy information 650, a pair status 651 shows whether the volume is mirrored or not, and its status if it is being mirrored. The PAIR status in Fig. 6 indicates that the primary and secondary volumes are mirrored.

The remote storage system information includes a serial 652, which indicates the product serial number of the paired storage system, and a location 653, which indicates the location of the paired storage system. When a user clicks the triangle button corresponding to the serial 652, information about the available storage systems connected to the local storage system described in the storage system information 630 is shown.

The port information includes a path group 654, which shows all path groups defined to the local storage system, and their status. If a path group does not connect to the selected remote storage system, then the status shows "N/A." If it is connected and available to use, then the status shows "RDY." The order from top to

bottom implies priority for use of the path group. For example, in Fig. 6, the path group "T3 up to 5MB/s" has the higher priority than "Internet," and the primary storage system 100a selects "T3 up to 5MB/s" when transferring data to the secondary storage system 100b. A user can change this order using this user interface.

When a user selects one of path group from the path group 654, then information for the selected path group appears in a change path group name 655, a status field 656, and a constraints field 657. A user can input a new name into the change path group name 655 in order to change the name. The status field 656 shows detailed status for the selected path group. The status can be one of the statuses of "available," "unavailable," or "temporarily unavailable," which have been described herein above. Note that if a user selects "unavailable," then the primary storage system 100a does not use the path group for transferring the data on the volume.

Many kinds of constraints for the selected path group can appear in the constraints field 657 in various embodiments of the present invention. When a user clicks the triangle button corresponding the constraints field 657, the constraints for the selected path group are displayed. If a constraint is applied to the path group, a check mark is shown on the left of the constraint, as shown in Fig. 6. If a user selects one of the constraints shown, then an appropriate window appears (not shown in Fig. 6). For example, if the user selects the "TP up to 5 MB/s" constraint shown in Fig. 6, then a window 700 illustrated by Fig. 7 is presented to the user. Using the dialog in the window 700, the user can input necessary information to set a throughput constraint, for example. When the user clicks an apply button 750, then the constraint information set up by the user is read and applied. Clicking a clear button 760 clears the current constraint information, causing the check mark icon in the constraints field 657 to disappear. The user fills in the necessary information using the management window 600 in Fig. 6. When, the user clicks an apply button 660, the constraint information input by the user is read and applied. The information is applied by the primary storage system 100a, which either creates or changes the path selection table 300 and path group table 400, according to the constraint information entered by the user. Further, the management console 120a maps a path group name entered by the user in the path group field 654 into a set of port numbers, and translates the path group name to the port numbers. For example, the path group "T3 up to 5MB/s" is translated to a port 0 and a port 1. Then, the management

console 120a sends the port numbers along with a volume number and constraints to the primary storage system 100a.

Implementation Examples

The present invention will next be described with reference to examples of using some of the various functions and features of various specific embodiments thereof. This section is intended to be merely illustrative of some of the many ways that specific embodiments of the present invention can use constraints as described herein above. Note that these examples use only two networks of differing types, such as an expensive network and an inexpensive network, for clarity of explanation. However, as is apparent to those skilled in the art, many different configurations may be readily prepared using a variety of network types in accordance with various specific embodiments of the present invention.

Fig. 8 illustrates a flowchart of representative processing in an implementation that uses an expensive network below a particular throughput or busy rate in a specific embodiment of the present invention. In the example implementation shown in Fig. 8, an expensive network is used if throughput or busy rate is below a maximum throughput or busy rate. When a user sets the "throughput constraint" for an expensive network as described herein above, the use of the expensive network is kept below the maximum throughput. Further, if the user sets the "busy rate constraint" for the networks, then he can use the networks below the maximum busy rate. This example implementation is representative of a situation in which users are allowed to use expensive networks under a certain data throughput. When the maximum throughput is exceeded, the users may incur additional charges, or network performance may significantly degrade.

The flowchart in Fig. 8 shows the constraint strategy which a user configures using the management console 120a in order to cause the primary storage system 100a to use an expensive network below a maximum throughput or busy rate, but use an inexpensive network for traffic if the throughput or busy rate exceeds the maximum specified in the constraint. In a step 800, using the user interface described in Fig. 6 and Fig. 7, the user makes the expensive network available up to predetermined maximum throughput, and gives the expensive network the first priority. Then, in a step 810, again using the user interface described in Figs. 6 and 7, the user makes the

inexpensive network available without constraint, and gives the inexpensive network the second priority. After the user has configured the constraint strategy according to the above steps, the primary storage system 100a transfers data to the secondary storage system 100b, according to the flowchart in Fig. 5. As previously described herein above, and with reference to Fig. 5, the primary storage system 100a selects an expensive network for sending traffic until the preset maximum throughput is reached. Once the maximum throughput is reached, the primary storage system 100a selects the inexpensive network since the expensive network no longer satisfies the constraint. Similarly, the user can set a busy rate constraint in step 800, as well.

Fig. 9 illustrates a flowchart of representative processing in an implementation that uses an inexpensive network during night operations in a specific embodiment of the present invention. In the example implementation shown in Fig. 9, an inexpensive, public network is used during nighttime operations. This example implementation is representative of a situation in which users are allowed to use public networks during nighttime, but avoid daytime public network access. Because public networks tend to have high traffic in the daytime, and transferring remote copy data through the public networks affects other services, like e-mails and web access, the user restricts use of the public network only to nighttime operations. In order to avoid using the public network during daytime operations, the user sets a "time constraint" for the public network. For example, the user may set a time constraint of "9:00 am to 9:00 pm" in order to prohibit the primary storage system 100a from using the public network.

The flowchart in Fig. 9 shows the constraint strategy which a user configures using the management console 120a in order to cause the primary storage system 100a to use inexpensive networks during nighttime. In a step 900, using the user interface described in Fig. 6 and Fig. 7, the user makes the inexpensive networks available only for nighttime (e.g. 9:00 pm to 6:00 am) use, and gives the inexpensive networks first priority. Then, in a step 910, again using the user interface described in Figs. 6 and 7, the user makes the inexpensive networks available without constraint, and gives the inexpensive networks second priority. After the user has configured the constraint strategy according to the above steps, the primary storage system 100a transfers data to the secondary storage system 100b, according to the flowchart in Fig. 5. As previously described herein above, and with reference to Fig. 5, the primary storage system 100a selects an inexpensive network for sending traffic from the time period

during 9:00 pm to 6:00 am. At other times, the primary storage system 100a selects the expensive network since the inexpensive network no longer satisfies the constraint. Similarly, the user can set a busy rate constraint in step 900, as well.

Fig. 10 illustrates a drawing of a representative system configuration in another specific embodiment of the present invention. In the example implementation shown in Fig. 10, the primary storage system 100a uses an inexpensive network except in case of an emergency. This example implementation is representative of a situation in which users subscribe to expensive networks on a pay per use basis. There are many different types of emergency cases that may be detected and responded to in various specific embodiments of the present invention. A brief sample of representative emergency cases will be described here. For example:

(1) When the inexpensive networks have high traffic and the primary storage system has a great deal of pending data. When an external network monitor that monitors traffic over the networks observes high traffic in the inexpensive network, the network monitor notifies the primary storage system 100a. Then the primary storage system 100a diverts traffic to other networks until the external network monitor indicates that the traffic in the inexpensive network has diminished.

(2) When the primary storage system has too much pending data. Generally, an inexpensive network is slower than an expensive network. So, data to be transferred to the secondary storage system accumulates in the primary storage system until there is sufficient network bandwidth available to move the accumulated data to the secondary storage system. If the inexpensive network continues to be slow, the primary storage system can, upon detecting this condition, switch to using a more expensive, and faster, network to send the accumulated data to the secondary storage system. In order to avoid a situation where the accumulated data makes it no longer possible to maintain a mirror image copy of the primary storage system data at the secondary storage system, the primary storage system monitors how much pending data has accumulated, and uses the more expensive, and faster, networks when the accumulated data exceeds a threshold.

(3) When errors exceed a threshold. The primary storage system monitors how many errors have occurred in transferring data through the networks and calculates an error count, which may be a percentage, for example. The primary storage system switches to a more expensive network when the error count for the inexpensive network exceeds a threshold. The threshold may be provided by a customer. While using

expensive networks, the primary storage system 100a sends some data over the inexpensive network at regular intervals, to perform a trial data transfer. The primary storage system 100a selects the remaining path groups to perform non-trial data transfers. The primary storage system 100a ceases using the expensive networks if the error count for the inexpensive networks falls below the threshold. This technique is useful in specific embodiments in which a TCP/IP protocol network is used as the inexpensive network transferring protocol, because a high degree of errors in such TCP/IP networks often indicates a high volume of traffic in the network.

(4) When errors occur. The primary storage system monitors for the presence of errors that occur in transferring data through the networks. The primary storage system uses an expensive network as an alternate path for an inexpensive network, and switches to the expensive network when an error is detected in the inexpensive network. This technique is useful in specific embodiments in which the primary storage system first attempts to transfer data via the inexpensive network. If this fails, the primary storage system uses the expensive network.

As shown in Fig. 10, two storage systems, the primary storage system 100a and the secondary storage system 100b, comprise one configuration for using a remote storage backup system. A network monitor 1000, connects to network 140a and network 140b, and management console 120a. The network monitor 1000 monitors activity in networks 140a and 140b. A path 1020 connects the network monitor 1000 to networks 140a and network 140b. A path 1010 connects the network monitor 1000 to the management console 120a. A path 1020 and a path 1010 may be parts of the same network, such as the Internet, for example. If the network monitor 1000 detects a high traffic volume in network 140a, then the network monitor 1000 sends a message to the management console 120a.

Fig. 11 illustrates a diagram of a representative network monitor message in another specific embodiment of the present invention. In the representative message format illustrated by Fig. 11, a network name 1100 corresponds to the network name registered in the path group 654 in Fig. 6. For example, a network name of "T3 up to 5MB/s" or "Internet" can be used. One purpose for the network name 1100 is to make the network identifiable by the management console 120a. A warning 1110 shows a type of warning that the network monitor 1000 discovered while monitoring the network. A variety of different types of warnings can be used in various specific embodiments of the

present invention. For example, in a specific embodiment, warnings for "Overload" and "Change to Normal" are provided. An "Overload" warning indicates that the network monitor 1000 found an overload condition within the network being monitored. An "Overload" warning includes a current busy rate that the network monitor 1000 determined during monitoring the network. A "Change to Normal" warning indicates that the network monitor 1000 found a network 140a has returned to a traffic volume level lower than a threshold busy rate. A current date and time field 1120 indicates a time when the network monitor 1000 issued the message to the management console 120a.

Fig. 12 illustrates a flowchart of representative processing in an implementation that uses network monitor in a specific embodiment of the present invention. In the example implementation shown in Fig. 12, the network monitor 1000 performs steps 1200 to 1230, and the management console 120a performs steps 1240 to 1260. In a step 1200, the network monitor 1000 monitors networks for a change in situation, such as a load change, an emergency, and the like. If a load situation change is detected by step 1200, then, in a decisional step 1210, a determination is made whether the change is from normal to overload. If the situation change is from normal to overload then, in a step 1220, the network monitor 1000 stores "Overload" into the warning field 1110 of a message having a format such as the message format described herein above with reference to Fig. 11, and sends the message to the management console 120a. Otherwise, if in step 1210 it is determined that the situation changed from overload to normal, then in a step 1230, the network monitor 1000 stores "Normal" into the warning field 1110 of the message, and sends the message to the management console 120a.

In a decisional step 1240, responsive to receiving the message sent by the network monitor 1000, the management console 120a checks the warning field 1110, to see if the warning field 1110 stores an "Overload" or a "Normal" condition type. If the warning field 1110 stores an "Overload," then, in a step 1250, the management console 120a sets the status 440 for the network to "temporarily unavailable" in the path group table 400. Otherwise, if in step 1240, it is determined that the warning 1110 stores an "Normal," then in a step 1260, the management console 120a sets the status 440 for the network to "available." As described herein above with reference to the flowchart in Fig. 5, the primary storage system 100a will avoid using a network having a status 440 of "temporarily unavailable" or "unavailable." Accordingly, the networks that the network

monitor 1000 determines are overloaded will not be selected by the primary storage system 100a.

Fig. 13 illustrates a flowchart of a representative processing in an implementation that uses an expensive network in emergency situations in a specific embodiment of the present invention. In the example implementation shown in Fig. 13, an inexpensive network is used if the workload situation is normal. When a user sets the error rate and outboard constraints for the inexpensive network as described herein, the use of the expensive network is reserved only for emergencies. This example implementation is representative of a situation in which users are allowed to use expensive networks only to deal with emergency situations. When the error rate constraint is exceeded, the network performance may be significantly degraded, causing the secondary storage system to be incapable of preserving a mirror image of the primary storage system. In a step 1300, the network monitor 1000 is configured to monitor inexpensive network 140a. A predetermined threshold for workload for the network 140a is configured using the user interface described above with reference to Figs. 6 and 7. Once configured, the network monitor 1000 performs the processing described above with reference to Fig. 12.

In a step 1310, using the user interface described in Fig. 6 and Fig. 7, the user makes the inexpensive network 140a available with an error rate constraint, such as a predetermined threshold for the error rate, for example, and an outboard constraint, and gives the inexpensive networks first priority. The outboard constraint causes inputs from the network monitor 1000 to be reflected to the path group table 400 in the primary storage system 100a, as described herein above with reference to Fig. 12. In a step 1320, the user makes the expensive network 140b available without a constraint, and gives the expensive network second priority. Using this constraint strategy, the primary storage system 100a selects the inexpensive network 140a, so long as the network monitor 1000 determines that there are no overloads or emergency conditions in the inexpensive network 140a. If an overload is detected by the network monitor 1000, this information is forwarded to the management console 120a, which reflects this condition in the status field 440 for the inexpensive network 140a in the path group table 400. A change to the status field 440, causes the primary storage system 100a to alter its selection of networks, by choosing the expensive network 120b until the overload situation in the inexpensive network 140a is relieved.

The preceding has been a description of the preferred embodiment of the invention. It will be appreciated that deviations and modifications can be made without departing from the scope of the invention, which is defined by the appended claims.

What is claimed is:

1. A storage system apparatus, comprising:

at least one of a plurality of disk drives;

a memory, operable to contain path selection information;

a plurality of ports, providing switch-able connection to a plurality of networks; and

a processor;

wherein said plurality of networks each has at least one of a plurality of user provided policies associated therewith, and wherein said processor, based upon monitoring of at least one of a plurality of conditions in said plurality of networks, selects at least one of said plurality of ports connecting said plurality of networks, based upon a comparison of said at least one of a plurality of conditions in said plurality of networks to a plurality of user provided policies.

2. The storage system apparatus of claim 1, wherein said at least one of a plurality of conditions comprises at least one of a throughput, a busy rate, an error rate, and a presence of an error.

3. The storage system apparatus of claim 1, further comprising a plurality of status indications, said plurality of networks each having at least one of said plurality of status indications associated therewith; and wherein said processor determines based upon said status indication whether to select a port from said at least one of a plurality of ports connecting said plurality of networks.

4. The storage system apparatus of claim 3, further comprising a network monitor, said network monitor operable to detect a condition within at least one of said plurality of networks, and thereupon set said value in said status indication.

5. The storage system apparatus of claim 3, wherein said status indication comprises at least one of available, temporarily unavailable, and unavailable.

6. The storage system apparatus of claim 1, wherein said policy comprises at least one of a threshold, a maximum, a minimum, an average, a mean, a limit, a constraint, a priority, and a target.

7. The storage system apparatus of claim 1, wherein said plurality of networks are grouped into a plurality of path groups, wherein said policies are associated with networks in a particular path group.

8. The storage system apparatus of claim 7, wherein said at least one of a plurality of disk drives comprises at least one of a plurality of volumes.

9. The storage system apparatus of claim 8, wherein each of said at least one of a plurality of volumes is permitted to access networks of at least one of said plurality of path groups.

10. A method for minimizing cost of network access by a storage apparatus, said method comprising:

specifying a first network to be used for transferring data;

specifying a constraint for said first network;

specifying a second network to be used for transferring data;

transferring data using said first network when conditions in said first network are in accordance with said constraint, otherwise transferring data using said second network.

11. The method of claim 10, further comprising:

transferring a portion of said data using said first network even when conditions in said first network are not in accordance with said constraint as a test;

monitoring conditions in said first network during said test; and

returning to transferring data using said first network when said test reveals that conditions in said first network are again in accordance with said constraint.

12. The method of claim 10, wherein said first network is relatively less expensive to use than said second network.

13. The method of claim 10, wherein specifying said constraint for said first network comprises specifying at least one of a throughput, a busy rate, an error rate, and a presence of an error.

14. The method of claim 10, wherein said first network is a public network and said second network is a private network.

15. The method of claim 10, further comprising:
making said first network a higher priority network than said second network.

16. The method of claim 10, further comprising:
detecting an abnormal condition in said first network and thereupon transferring data using said second network.

17. A method for selecting a network, said method comprising:
monitoring at least one of a plurality of conditions in a plurality of networks;
comparing said at least one of a plurality of conditions in said plurality of networks to at least one of a plurality of user provided policies; and
selecting at least one of a plurality of ports connecting to said plurality of networks;
wherein said plurality of networks each has at least one of said plurality of user provided policies associated therewith.

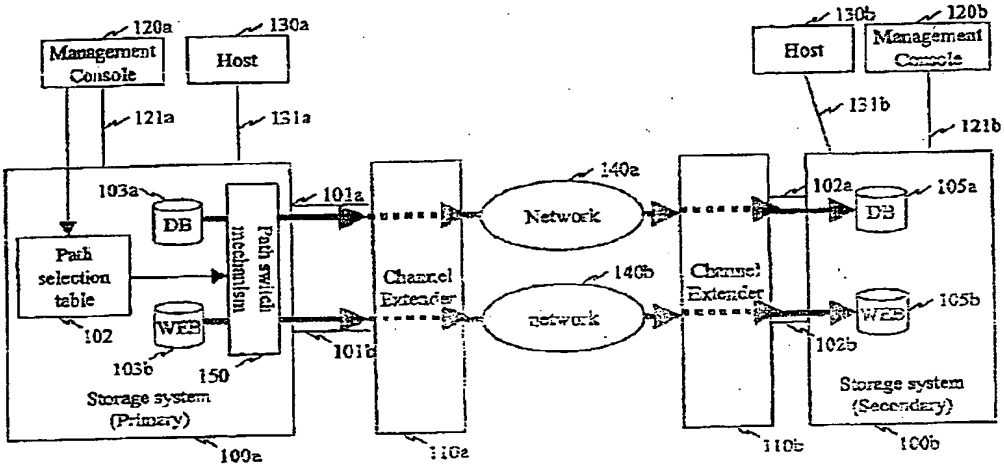
18. The method of claim 17, wherein selecting at least one of a plurality of ports connecting to said plurality of networks comprises:
determining based upon a status indication whether to select a port from said at least one of a plurality of ports connecting said plurality of networks.

19. The method of claim 17, further comprising:
associating said plurality of networks with a plurality of path groups;
wherein said at least one of a plurality of policies is associated with at least one of a plurality of path groups.

20. The method of claim 17, wherein monitoring at least one of a plurality of conditions in a plurality of networks comprises:
using a network monitor to detect a condition within at least one of said plurality of networks, and thereupon set a value in a status indication.

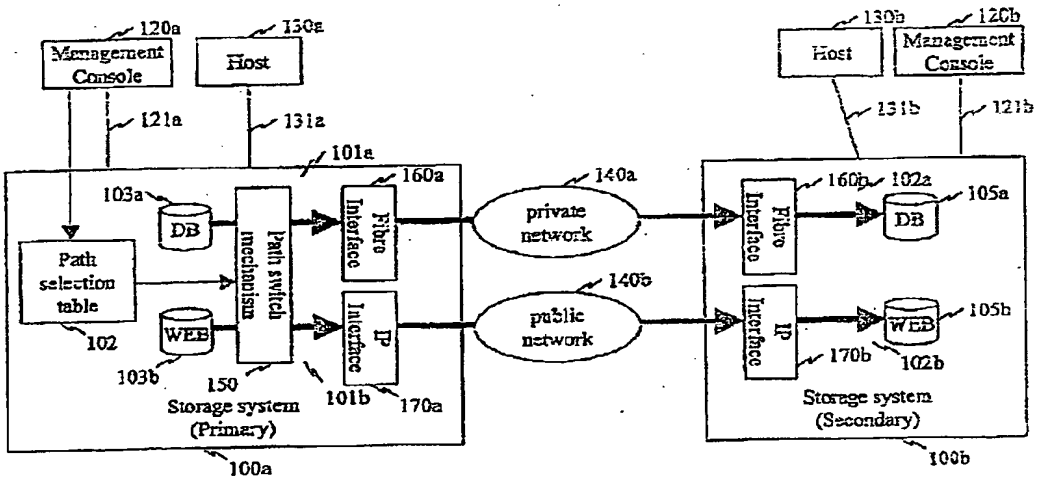
【図１Ａ】

Fig 1A



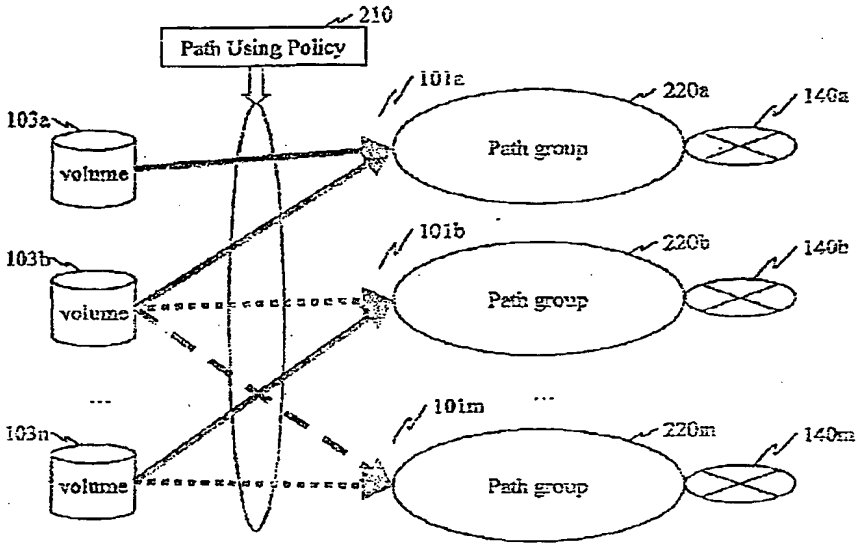
【図１Ｂ】

Fig 1B



【図２】

Fig 2



【図３】

Fig 3 path selection table 300

310 Volume number	320a Path group number	320b Path group number
0	0	NULL
1	1	0
2	0	NULL
3	1	NULL
4	1	0

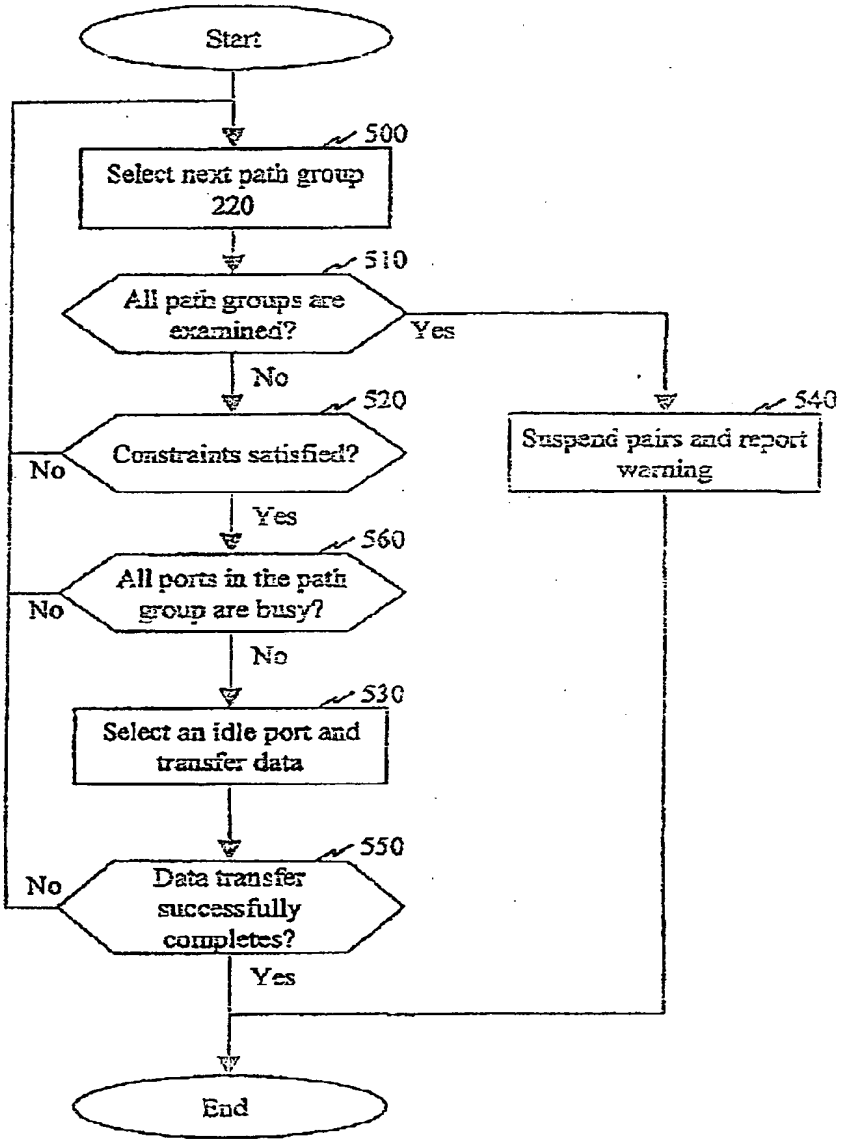
【図４】

Fig 4 path group table 400

410	420	440	430a	430b	430c
Path group number	Constraint	Status	Remote link number	Remote link number	Remote link number
0	Max 5MB/s	Unavailable	0	1	NULL
1	9:00pm to 6:00am only	Available	2	3	4
2	Busy rate less than 70%	Available	5	6	NULL
3	Error rate less than 5%	Temporaliy Unavailable	7	8	NULL
4	Outboard control	Temporaliy Unavailable	9	10	11

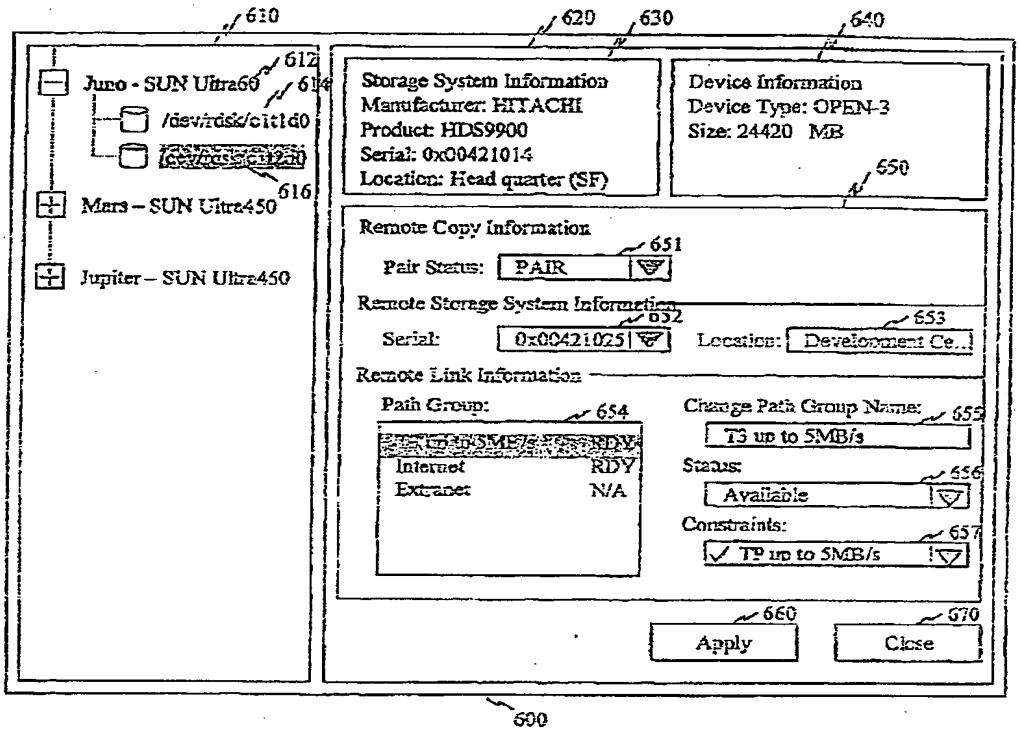
【図５】

Fig 5 path selecting algorithm



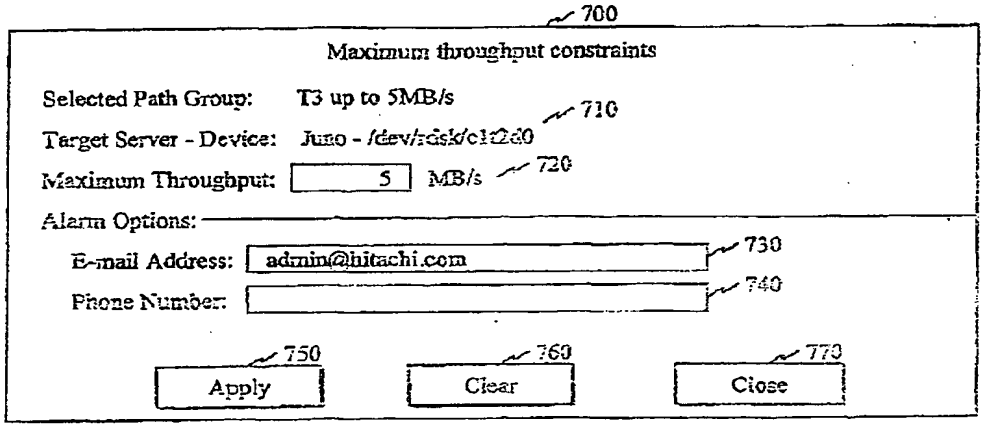
【図６】

Fig 6 User interface for setting path constraints



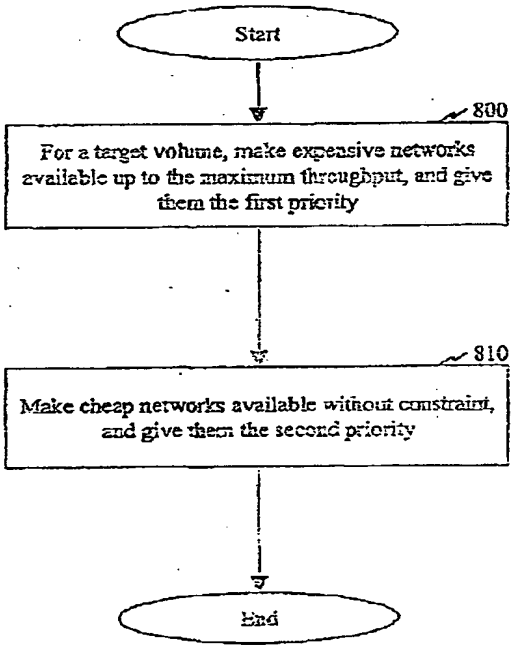
【図７】

Fig 7 Maximum throughput dialog



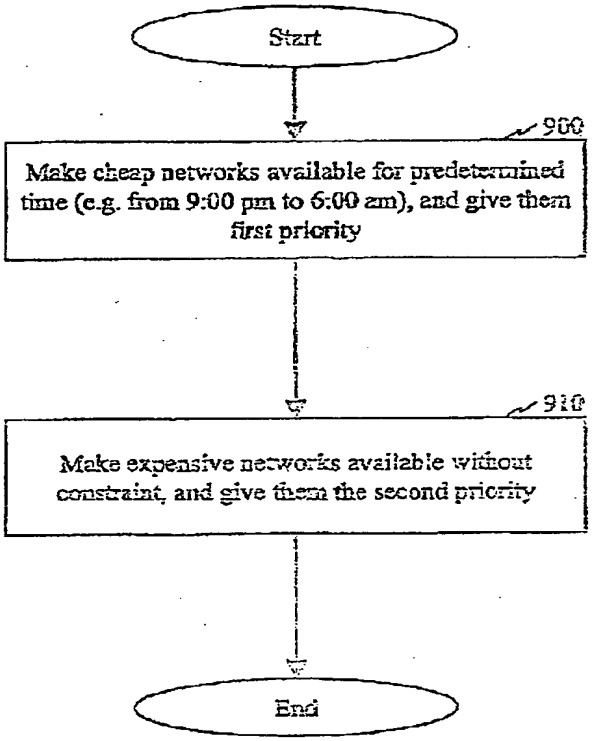
【図８】

Fig 8 Flow chart for using expensive networks below maximum throughput or busy rate



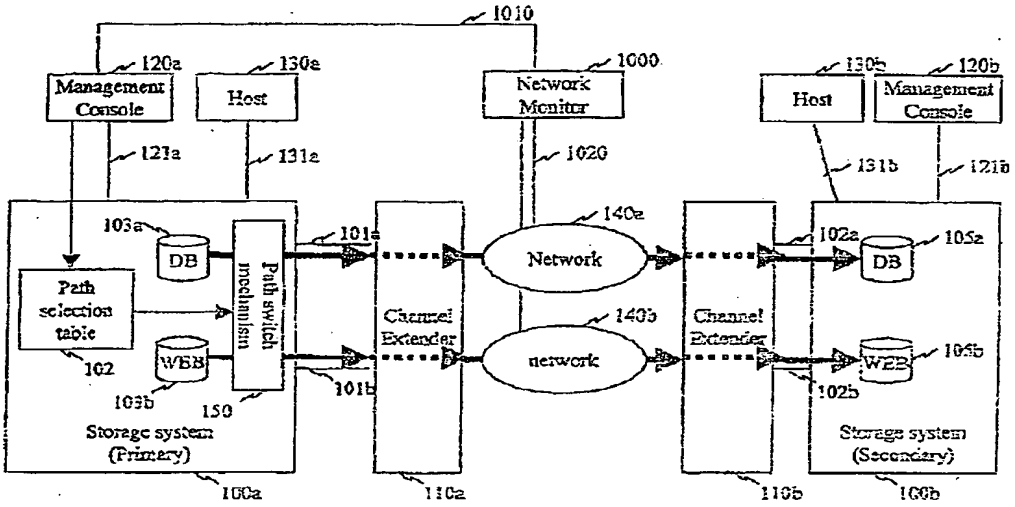
【図９】

Fig 9 Flow chart for using cheap networks only daytime



【図１０】

Fig 10 System Configuration with Network Monitor



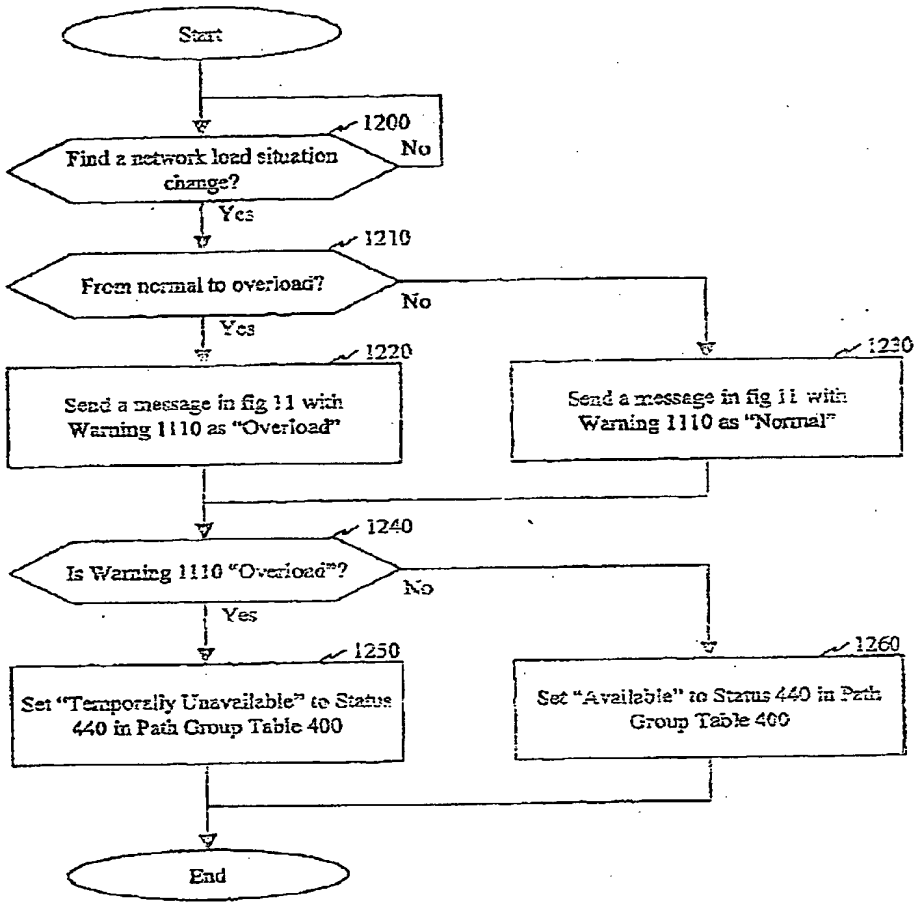
【図１１】

Fig 11 Information sending from Network Monitor to Management Console

Network Name	Internet	1100
Warning	Overload (80% busy rate)	1110
Current Date and Time	12-12-2000:13:46	1120

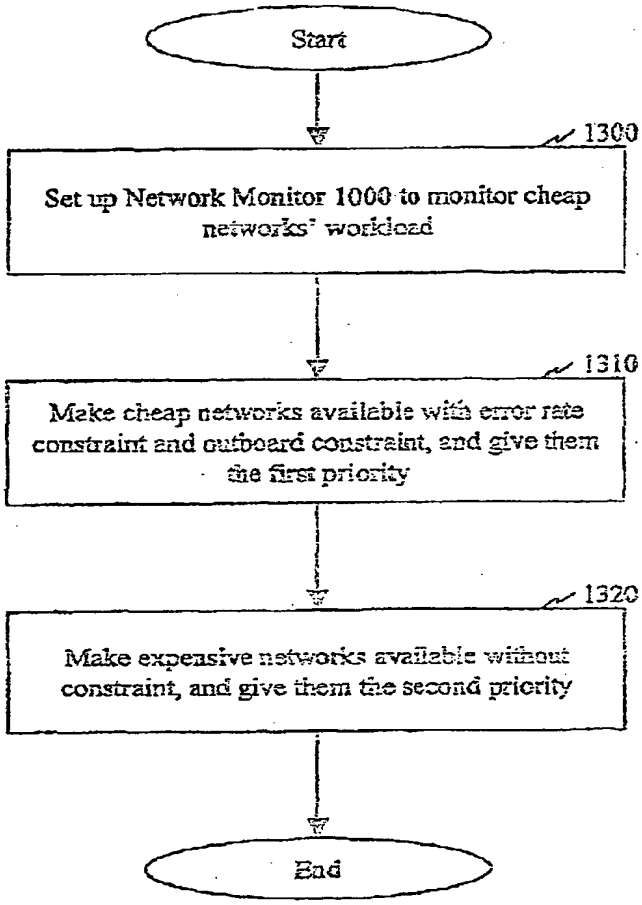
【図１２】

Fig 12 Flowchart for monitoring networks



[図１３]

Fig 13 Flow chart for using expensive network only in emergency



Path Selection Methods for Storage Based Remote Copy

ABSTRACT OF THE DISCLOSURE

The present invention provides techniques for managing data flow over a plurality of connections between primary and remote storage devices. In a representative example embodiment, when the primary storage system copies data to the secondary storage system, it chooses one of a plurality of networks connecting it to the secondary storage system, depending upon a users' policy. Since networks have different characteristics, in terms of, for example, performance, security, reliability, and costs, the user can specify which network(s) are used under various circumstances, i.e., daytime operation, nighttime operation, normal operation, emergency, and so forth. The storage systems comprise a mapping of volumes and ports. When performing copy operations, the primary storage system finds a volume storing the data, and available ports by accessing the mapping. The mappings are based upon policies that are input by a user.